Deep Reinforcement Learning for the Real Time Control of Stormwater Systems

Abhiram Mullapudi[a], Matthew J. Lewis[c], Cyndee L. Gruden[b], Branko Kerkez[a,*]

[a] *Department of Civil and Environmental Engineering, University of Michigan, Ann Arbor, MI, USA*

[b] *Department of Civil and Environmental Engineering, University of Toledo, Toledo, OH, USA*

[c] *Michigan Aerospace, Ann Arbor, MI, USA*

Abstract

A new generation of smart stormwater systems promises to reduce the need for new construction by enhancing the performance of the existing infrastructure through real time control. Smart stormwater systems dynamically adapt their response to individual storms by controlling distributed assets, such as valves, gates and pumps. This paper introduces a real time control approach based on Reinforcement Learning (RL), which has emerged as a state-of-the-art methodology for autonomous control in the artificial intelligence community. Using a deep neural network, an RL based controller learns a control strategy by interacting with the system it controls - effectively trying various control strategies until converging on those that achieve a desired objective. This paper formulates and implements an RL algorithm for the real time control of urban stormwater systems. This algorithm trains an RL agent to control valves in a distributed stormwater system across thousands of simulated storm scenarios,

*Corresponding author

seeking to achieve water level and flow setpoints in the system. The algorithm is first evaluated for the control of an individual stormwater basin, after which it is adapted to the control of multiple basins in a larger watershed ($4km^2$). The results indicate that RL can very effectively control individual sites. Performance is highly sensitive to the reward formulation of the RL agent. Generally, more explicit guidance — encoded as a more complex reward formulation – leads to better control performance, and more rapid and stable convergence of learning process. While the control of multiple distributed sites also shows promise in reducing flooding and peak flows, the complexity of controlling larger systems comes with a number of caveats. The RL controller's performance is very sensitive to the formulation of the deep neural network and requires a significant amount of computational resource to achieve a reasonable performance enhancement. Overall, the controlled system significantly outperforms the uncontrolled system, especially across storms of high intensity and duration. A frank discussion is provided, which should allow the benefits and drawbacks of RL to be considered when implementing it for the real time control of stormwater systems. An open source implementation of the full simulation environment and control algorithms is also provided.

Keywords: real time control, reinforcement learning, smart stormwater systems

1. Introduction

Urban stormwater and sewer systems are being stressed beyond their intended design. The resulting symptoms manifest themselves in frequent flash floods Laris Karklis and Muyskens (2017) and poor receiving water quality Watson et al. (2016). Presently, the primary solution to these challenges is the construction of new infrastructure, such as bigger pipes, basins, wetlands, and other distributed storage assets. Redesigning and rebuilding the existing stormwater infrastructure to keep in pace with the evolving inputs is cost prohibitive for most communities Kerkez et al. (2016). Furthermore, infrastructure is often upgraded on a site-by-site basis and rarely optimized for system-scale performance. Present approaches rely heavily on the assumption that these individual upgrades will add up to cumulative benefits, while the contrary has actually been illustrated by studies evaluating system-level outcomes Emerson et al. (2005). The changing and highly variable nature of weather and urban environments demands stormwater solutions that can more rapidly adapt to changing community needs.

Instead of relying on new construction, a new generation of smart stormwater systems promises to dynamically re-purpose existing stormwater systems. These systems will use streaming sensor data to infer real time state of a watershed and respond via real time control of distributed control assets, such as valves, gates and pumps Kerkez et al. (2016). By achieving system-level coordination between many distributed control points, the size of infrastructure needed to reduce flooding and improve water quality will become smaller. This presents a non-trivial control challenge, however, as any automated decisions must be carried with regard to public safety and must account for the physical complexity inherent to urban watersheds Mullapudi et al. (2017); Schütze et al. (2004).

In this paper, we investigate *Deep Reinforcement Learning* for the real time control of stormwater systems. This approach builds on very recent advances in the artificial intelligence community, which have primarily focused on the control of complex autonomous systems, such as robots and autonomous vehicles Mnih et al. (2015); Lillicrap et al. (2015). In this novel formulation, our algorithm will *learn* the best real time control strategy for a distributed stormwater system by efficiently quantifying the space of all possible control actions. In simple terms, the algorithm attempts various control actions until discovering those that have the desired outcomes. While such an approach has shown promise across many other domains, it is presently unclear how it will perform and scale when used for the real time control of water systems, specifically urban drainage networks.

The fundamental contribution of this paper is a formulation of a control algorithm for urban drainage systems based on Reinforcement Learning. Given the risk to property and public safety, it is imprudent to hand over the control of a real-world watershed to a computer that learns by mistake. As such, a secondary contribution is the evaluation of the Reinforcement Learning algorithm across a series of simulations, which span various drainage system complexities and storms. The results will illustrate the benefits, limitations, and requirement of Reinforcement Learning when applied to urban stormwater systems. To our knowledge, this is the first formulation of Deep Reinforcement Learning for the control of stormwater systems. The results of this study stand to support a foundation for future studies on the role of Artificial Intelligence in the control of urban water systems.

1.1. ***Real time control of urban drainage systems***

Since the European Union's Directive on water policy The European Parliament and the council of European Union (2000), there has been a significant push towards the adoption of real time control for improving wastewater and sewer systems Schütze et al. (2004); Mollerup et al. (2016). During the past decade, Model Predictive Control (MPC) has emerged as a state-of-the-art methodology for controlling urban drainage and sewer networks. MPC has been used to regulate dissolved oxygen in the flows to aquatic bodies Mahmoodian et al. (2017), control inflows to wastewater treatment plants Pleau et al. (2005), and enhance the system-level performance and coordination of sewer network assets Mollerup et al. (2016); Meneses et al. (2018). These and many other Wong and Kerkez (2018) applications have illustrated the benefits of control, the biggest of which is the ability to cost-effectively re-purpose existing assets in real time without the need to build more passive infrastructure.

The performance of MPC depends on the extent to which the underlying process can be approximated using a linear model Van Overloop (2006). A benefit of this linearity assumption is the ability to analytically evaluate the stability, robustness and convergence properties of the controller Ogata (2011), which is valuable when providing safety and performance guarantees. Network dynamics of storm and sewer systems and transformations of the pollutants in runoff are known to be heavily non-linear, however. This demands a number of approximations and a high level of expertise when applying MPC. Furthermore, real-world urban watersheds are prone to experiencing pipes blockages, sensor breakdowns, valve failures, or other adverse conditions. Adapting and re-formulating linear control models to such non-linear conditions is difficult, but is being addressed by promising research Vezzaro and Grum (2014). The constraints of linear approximations and the need for adaptive control algorithms open the door to exploring other control methodologies, such as the one presented in this paper.

2. Reinforcement Learning

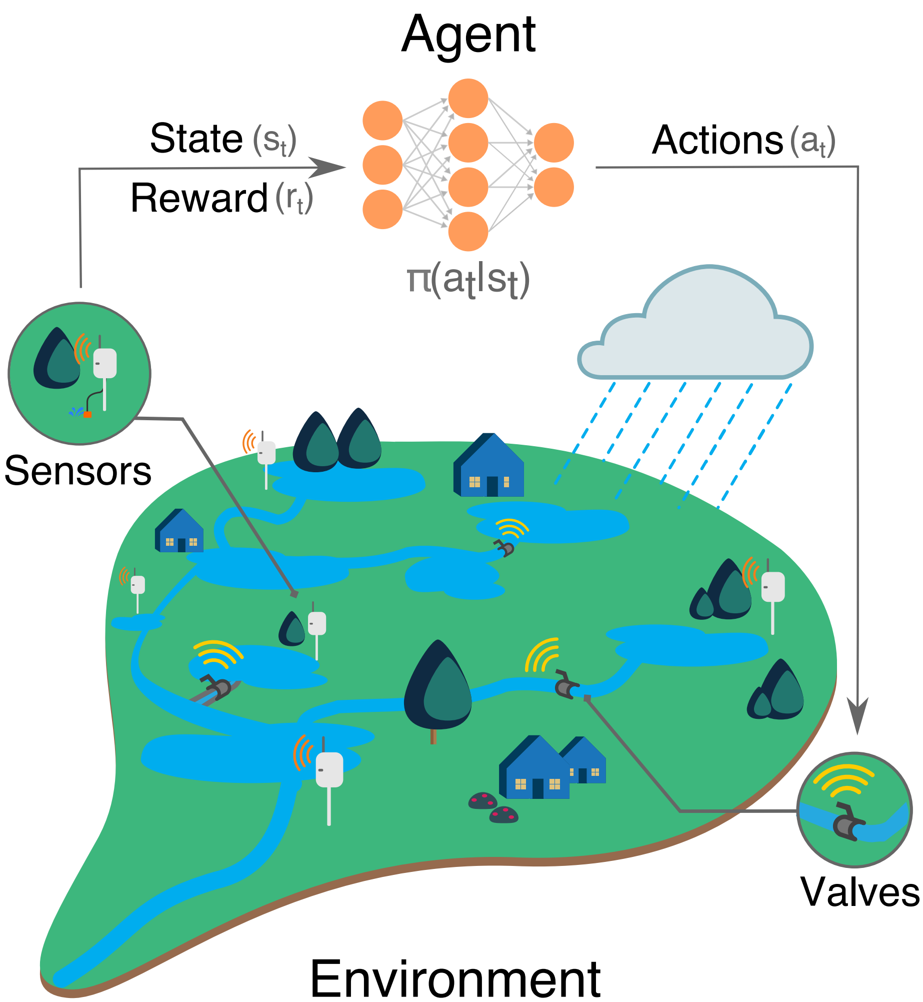

Figure 1: During a storm event, a reinforcement learning controller observes the state (e.g. water levels, flows) of the stormwater network and coordinates the actions of the distributed control assets in real time to achieve watershed-scale benefits.

Across the artificial intelligence and behavioral research communities, Reinforcement Learning (RL) has emerged as a state-of-the-art methodology for autonomous control and planning systems. Unlike in classical feedback control, where the controller carries out a pre-tuned and analytical control action, an

RL controller (i.e. an RL agent) learns a control strategy by interacting with the system - effectively trying various control strategies until learning those that work well. Rather than just learning one particular control strategy, an RL agent continuously attempts to improve its control strategy by assimilating new information and evaluating new control strategies Sutton and Barto (1998). RL can be used in a model free context since the system's dynamics are implicitly learned by evaluating various control actions. Leveraging the recent advancements in Deep Neural Networks and the computational power afforded by the high performance clusters (HPCs), RL agents have been able plan complex tasks, such as observing pixels to play video games at a human level Mnih et al. (2015), defeating world champions in the game of GO Silver et al. (2017b), achieving "superhuman" performance in chess Silver et al. (2017a), controlling high speed robots Kober et al. (2013), and navigating autonomous vehicles Ng et al. (2006). Despite the wide adoption of Deep Neural Network based Reinforcement Learning (Deep RL) in various disciplines of engineering, its adoption in civil engineering disciplines has been limited Abdulhai and Kattan (2003); Bhattacharya et al. (2003); Castelletti et al. (2010). Deep RL control has yet to be applied to the real time control of urban drainage systems.

Deep RL agents approximate underlying system dynamics implicitly, hence not requiring a simplified or linearized control model Sutton and Barto (1998). A Deep RL agent instantaneously identifies a control action by observing the network dynamic, thus reducing delay in the decision process Mnih et al. (2015); Silver et al. (2017a). The explorative nature of the Deep RL agents also enables the methodology to adapt its control strategy to changing conditions of the system Sutton and Barto (1998). Hence, Reinforcement Learning shows promise as a potential alternative or supplement to existing control methods for water systems. To that end, the goal of this paper is to formulate and evaluate of

Reinforcement Learning for the real time control of urban drainage systems. The specific contributions of the paper are:

1. The formulation and implementation of a reinforcement learning algorithm for the control of urban stormwater systems.
2. An evaluation of the control algorithm under a range of storm inputs and network complexities (single stormwater basins and an entire network), as well as an equivalence analysis that compares the approach to passive infrastructure solutions.
3. A fully open-sourced implementation of the control algorithm to promote transparency and permit for the direct application of the methods to other systems, shared on open-storm.org.

Symbol	Definition
$\boldsymbol{s_t}$	state observed by agent at time t
$\boldsymbol{a_t}$	action taken by agent at time t
$\boldsymbol{r_t}$	reward received by the agent at time t
$\boldsymbol{\pi}$	policy of the agent
$\boldsymbol{v(s_t)}$	value estimate for a given state s_t
$\boldsymbol{q(s_t, a_t)}$	action value estimate for a given state action pair s_t, a_t
$\boldsymbol{q}$	action value estimator
$\boldsymbol{q^*}$	target estimator
$\boldsymbol{\epsilon}$	rate of exploration
$\boldsymbol{\alpha}$	step size
$\boldsymbol{\gamma}$	discount factor
$\boldsymbol{h_t}$	basin's water level at time t
$\boldsymbol{f_t}$	channel's flow at time t
$\boldsymbol{H}$	desired water height in basin
$\boldsymbol{H_{max}}$	height threshold for flooding
$\boldsymbol{F}$	flow threshold for erosion

Table 1: Summary of notation used in paper.

3. Methods

3.1. *Reinforcement learning for stormwater systems*

When formulated as a Reinforcement Learning (RL) problem, the control of stormwater systems can be fully described by an agent and environment (Figure 1). The environment represents an urban stormwater system and the agent represents the entity controlling the system. At any given time t, the agent takes a control action a_t (e.g. opening a valve or turning on a pump) by observing any number of states s_t (e.g. water levels or flows) in the environment. Based on the outcomes of its action, the agent receives a reward r_t from the environment. The reward is formulated to reflect the specific control objectives. For example, an agent could receive positive reward for a preventing flooding or a negative reward for causing flooding. By quantifying these rewards in response to various actions over time, the agent learns the control strategy that will achieve its desired objective Sutton and Barto (1998). The agent's control actions in any given state are governed by its policy π. Formally, the policy is a mapping from a given state to the agent's actions:

$$\pi : s_t(\mathbb{R}^n) \rightarrow a_t(\mathbb{R}) \tag{1}$$

The primary objective of the RL control problem is to learn a policy that maximizes the total reward earned by the agent.

While the reward r_t at the end of each control action teaches the agent the immediate desirability of taking a particular action for a given state, it does not necessarily covey any information about the long-term desirability of that action. For many water systems, maximizing short-term rewards will not necessarily lead to the best long-term outcomes. An agent controlling a watershed or stormwater system should have the ability to take individual actions in the context of the entire storm duration. For example, holding water in a detention basin may initially provide high rewards since it reduces downstream flooding, but may lead to upstream flooding if a storm becomes too large. Instead of choosing an action that maximizes the reward r_t at time t, the agent seeks to maximize the expected long-term reward described by state-value v or action-value q.

$$v(s_t) = \mathbb{E}\left[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \dots \middle| s_t\right] = \mathbb{E}\left[\sum_{k=0}^{\infty}\left[\gamma^k r_{t+k+1} \middle| s_t\right]\right] \tag{2}$$

$$q(s_t, a_t) = \mathbb{E}\left[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \dots \middle| s_t, a_t\right] = \mathbb{E}\left[\sum_{k=0}^{\infty}\left[\gamma^k r_{t+k+1} \middle| s_t, a_t\right]\right] \tag{3}$$

The state-value provides an estimate for an instantaneous action, as well as potential future rewards that may arise after state s_t, discounted with a factor $\gamma (0 \leq \gamma \leq 1)$. The action-value provides a similar estimate conditioned, also however, on taking an action a_t in state s_t. The discount factor γ governs the temporal context of the reward. For example, a γ of 0 forces the agent to maximize the instantaneous reward, while a γ of 1 forces it to equally weigh all the rewards it might receive for present and future outcomes. γ is specific to the system being controlled and can vary based on the control objective Sutton and Barto (1998).

An RL agent can learn to control a system either by learning the policy directly Sutton et al. (2000). Alternatively, the agent can learn the state-value or action-value estimates and follow a policy that guides it towards the states with high estimates Sutton and Barto (1998). Several methods based on dynamic programming Watkins and Dayan (1992); Sutton (1991) and Monte Carlo sampling Sutton and Barto (1998) have been developed to learn the functions that estimate the policy and value functions. While these algorithms were computationally efficient and provided guarantees on the convergence, their application was limited to simple systems whose state action space can be approximated using lookup tables and linear functions Sutton and Barto (1998); Mnih et al. (2013).

Given the scale and the complexity of urban watersheds and stormwater networks, a simple lookup table or a linear function cannot effectively approximate the policy or value functions for each state the agent may encounter while controlling the system. As a simple example, considering just ten valves in a stormwater system and assuming that each valve has ten possible control actions (closed, 10% open, 20% open, ...) this gives 10^{10} (10 billion) possible actions that can be taken at any given state, making it computationally impossible to build an explicit lookup table for all possible states. This, however, is where very recent advances in Deep Learning, become important. It has been shown that for systems with a large state-action spaces, such as stormwater systems, these functions can be approximated by a deep neural network Sutton and Barto (1998); Mnih et al. (2015).

Deep neural networks are a class of feed forward artificial neural networks with large layers of inter connected neurons. This deeply layered structure permits the network to approximate highly complex functions Hornik et al. (1989), such as those needed for RL-based control. Each layer in the network generates its output by processing the weighted outputs from the previous layer. This means that each layer's output is more complex and abstract than its previous layer. Given the emergence of cheap and powerful computational hardware over the past decade – in particular graphical processing units (GPUs) and high performance clusters (HPCs) – deep neural networks and their variants have emerged as the state of the art in the approximation of complex functions in large state spaces LeCun et al. (2015a). This makes them a good candidate for approximating the complex dynamics across stormwater systems. For purposes of this paper, a brief mathematical summary of deep neural networks in provided in appendix A.

3.2. ***Deep Q Learning***

Deep reinforcement learning agents (deep RL) use deep neural networks as approximators for value or policy functions to control complex environments. In their relatively recent and seminal Deep Q Network(DQN) paper Mnih et al. (2015) Mnih et al. (2015) demonstrated the first such algorithm, which used deep neural networks to train an deep RL agent to play *Atari* video games at a human level. This algorithm identifies the optimal control strategy for achieving an objective by learning a function that estimates the action values or q-values. This function (i.e. q-function), maps a given state-action pair (s_t, a_t) pair to the action value estimate.

At the beginning of the control problem, the agent does not know its environment. This is reflected by assigning random q-values for all state-action pairs. Over time, as the agent takes actions, new information obtained from the environment is used to update these initial random estimates. After each action the reward obtained from the environment is used to incorporate the new knowledge:

$$q(s_t, a_t) \leftarrow q(s_t, a_t) + \alpha \left[r_{t+1} + \gamma \max_a q(s_{t+1}, a) - q(s_t, a_t) \right] \tag{4}$$

The more actions an agent takes at any given state, the closer it gets to converging to the true action value function Sutton and Barto (1998). The α (step-size) parameter governs how much weight is placed on the new knowledge Sutton and Barto (1998).

An agent will choose an action that maximizes its long-term reward. This process is known as exploitation since it greedily seeks to maximize a known long-term reward. This may not always be the best choice, however, since taking another action may lead the agent to discover a potentially better action, which it has not yet tried. As such, the agent also needs to explore its environment. This is accomplished by taking a random action periodically, just in case this action leads to better outcomes. In such a formulation, the *exploration vs. exploitation* is addressed via a ϵ-greedy policy, where the agent explores for ϵ percent of time and chooses an action associated with the highest action value for the rest. This gives the final policy for the RL agent:

$$\pi(s_t) = \begin{cases} \text{random } a, & \epsilon \\ \arg\max\limits_a q(s_t, a), & else \end{cases} \tag{5}$$

ϵ is often set at a high value (e.g. 50%) at the start of the learning process and gradually reduced to lower value (e.g. 1%) as the agent identifies a viable control strategy.

While there have been prior attempts to approximate the action value function using deep neural networks, they were met with minimal success since the learning is highly unstable Mnih et al. (2015). Mnih et al. (2015) Mnih et al. (2015) addressed this by introducing a replay buffer and an additional target neural network. The replay buffer acts as the RL agent's memory, which records only its most recent experience (e.g. the past 10^3 states transitions and rewards). During the training the RL agent randomly samples data from the replay buffer, computes the neural network's loss and updates its weights using stochastic gradient descent:

$$Loss = ||\big(r_t + \gamma \max_{a'} q^*(s_{t+1}, a')\big) - q(s_t, a_t)||^2 \tag{6}$$

This random sampling enables the training data to be uncorrelated and has been found to improve the training process. The target neural network q^* has the same network architecture as the main network q, but acts as a moving target to help stabilize the training process by reducing variance Mnih et al. (2015). Unlike the neural network approximating q, whose weights are constantly updated using gradient decent, q^* weights are updated sporadically (e.g. every 10^4 timesteps). For more background information, Mnih et al. (2015)Mnih et al. (2015) and Lillicrap et al. (2016)Lillicrap et al. (2015) provide an in-depth discussion on the importance of replay memory and target neural networks in training deep RL agents.

3.3. ***Evaluation***

Here, we investigate the real time control of urban stormwater infrastructure using Deep RL. To begin, we formulate and evaluate reward functions for the control of an individual stormwater basin. We then extend these lessons to the control of a larger, interconnected stormwater network. Given the relatively nascent nature of Deep RL, the need to account for public safety, and the desire to evaluate multiple control scenarios, a real-world evaluation is outside of the scope of this paper. As such, our analysis will be carried out in simulation as a stepping-stone toward real-world deployment in the future. To promote transparency and broader adoption, the entire source code, examples, and implementation details of our implementation are shared freely as an open source package[1].

3.4. *Study Area*

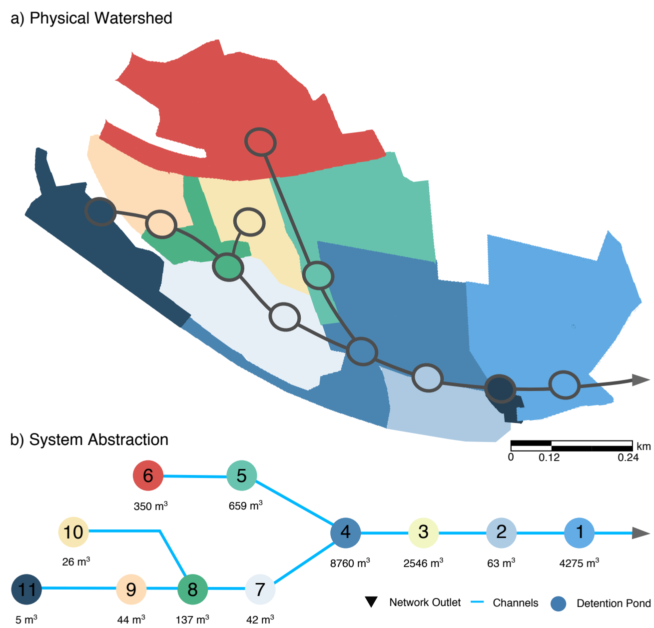

Figure 2: Stormwater system being controlled in this paper. The urban watershed includes a number of sub-catchments which drain to 11 stormwater basins. The first control scenario applies RL to the control of a single basin, while the second scenario evaluates control of multiple basins. Average volumes experienced by the ponds during a 25 year 6 hour storm event are presented.

Motivated by a real-world system, we apply RL control to a stormwater system inspired by an urban watershed Ann Arbor, Michigan, USA (2). Our

[1]https://github.com/kLabUM/rl-storm-control

choice to use this watershed is motivated by the fact that it has been retrofitted by our group with wireless sensors and control valves already Bartos et al. (2017) and will in the future serve as a real-world testbed for the ideas proposed in this paper. This headwater catchment features 11 interconnected stormwater basins that handle the runoff generated across $4km^2$ of predominantly urbanized and impervious sub-catchment areas. A Stormwater Management Model (SWMM) of the watershed has been developed and calibrated in prior, peer-reviewed studies Wong and Kerkez (2018). It is assumed that each controlled basin in the system is equipped with a $1m^2$ square gate valve. The valves can be partially opened or closed during the simulation, which represents the action taken by an RL agent. The states of the control problem are given by the water levels and outflows at each controlled location. Given the small size of the study area, as well as the need to constrain this initial study, uniform rainfall across the study area is assumed. Groundwater baseflow is assumed to be negligible, which has also been confirmed in prior studies Wong and Kerkez (2018).

3.5. ***Analysis***

Prior Deep RL studies have revealed that performance is dependent on the formulation of reward function, quality of neural networks approximating action value function, as well as the size of state space Sutton and Barto (1998); Henderson et al. (2017). This creates a number of "knobs", whose sensitivity must be evaluated before any conclusion can be reached regarding the ability to apply Deep RL to control real stormwater systems. As such, in this paper we formulate a series of experiments across two scenarios to characterize Deep RL's ability to control stormwater systems. In the first scenario, we control a single valve at the outlet of the watershed, comparing in particular performance under various reward function formulations. Given that Deep RL has not been used to control water systems, this will constrain the size of the state space to establish a baseline assessment of the methodology. In the second scenario, we scale these findings to simultaneously control multiple valves across the broader watershed and to analyze sensitivity to function approximation (neural networks). Finally, the system-scale scenario is subjected to storm inputs of varying intensities and durations to provide broader comparison of the benefits of the controlled system in relation to the uncontrolled system.

3.6. ***Scenario 1: Control of a single basin***

In this scenario, we train a deep RL agent to control the most downstream detention basin in the network (basin 1 in Figure 2). This basin was chosen because it experiences the total runoff generated in the watershed, and because its actions have direct impact on downstream water bodies. At any given point in time, the RL agent is permitted to set the basin's valve to a position between fully closed or open, in 1% increments (i.e. 0%, 1%, 2%, ..., 100% open) based on the water height in the basin. All other upstream basins remain uncontrolled.

The overall control objective is to keep the water height (state:$\{h_t\}$) in the basin below a flooding threshold H_{max} and the outflows from the basin (state :$\{f_t\}$) below a desired downstream flooding or stream erosion threshold F:

$$h_t \leq H_{max} \tag{7}$$

$$f_t \leq F \tag{8}$$

Three reward functions are formulated to reach this objective, varying in complexity from most simple to complex.

In the first reward function the RL agent receives a positive reward for maintaining the basin's outflow below the specified threshold, a negative reward for exceeding the threshold, as well as a larger but less likely negative reward if the basin overflows:

$$r_1(s_t) = \begin{cases} +1, & f_t \leq F \\ -1, & f_t > F \\ -10, & h_t > H_{max} \end{cases} \tag{9}$$

The reward function is represented visually in the first row of Figure 3. This reward function formulation is inspired from the classic inverted pendulum problem Watkins and Dayan (1992) where the agent receives +1 for success and -1 for failure.

The second reward function is formulated to exhibit a more complex and gradual reward structure. In lieu of a jagged or discontinuous "plus minus" reward structure, the agent is rewarded for reaching flows that are close to the desired flow threshold. It has been shown that more smooth and continuous rewards such as this may help the agent converge onto a solution faster Sutton and Barto (1998); Aytar et al. (2018). Visually, the reward function looks like a parabola (Figure 3), where the maximum reward is achieved when the flow threshold is met exactly:

$$r_2(s_t) = c_1(f_t - c_2)(f_t - c_3) \tag{10}$$

c_1,c_2,and,c_3 are constants representing the scaling and inflection points of the parabola. Here we choose c_1=-400 e, c_2=0.05, and c_3=0.15 to maintain the general scale of the first reward function. Note that this formulation does not explicitly include the local constraint on the basin's water level since the agent gets implicitly penalized by receiving a negative reward for low outflows.

The third reward function seeks to provide the most explicit guidance to the RL agent by embedding the most relative amount of information (third column, Figure 3). In this heuristic formulation, the agent receives the highest reward for keeping the basin empty (water levels and flows equal to zero). Intuitively, this reward formulation seeks to drain all of the water from the basin as fast as possible without exceeding the flow and height thresholds. If water level in the pond rises, the agent gets penalized, thus forcing it to release water. If flows remain below the flow threshold F, the agent is penalized linearly proportional to the water level in the basin, with a more severe factor applied if the height of the basin exceeds the height threshold H. If the outflow exceeds the flow threshold F an even more severe penalty is incurred:

$$r_3(s_t) = \begin{cases} c_1 - c_2 h_t, & h_t < H f_t \leq F \\ c_1 - c_3 h_t, & h_t \geq H f_t \leq F \\ -c_4 f_t - c_2 h_t + c_5, & h_t < H f_t > F \\ -c_4 f_t - c_3 h_t + c_5, & h_t \geq H f_t > F \end{cases} \tag{11}$$

The penalty rates are governed by a set of five parameters c=$\{c_1, c_2, c_3, c_4, c_5\}$, which were paramterized $\{2.0, 0.25, 1.5, 10, 3\}$ to match the scales of the other two reward functions.

3.7. Scenario 2: Controlling multiple basins

This scenario evaluates the ability of an agent to control multiple distributed stormwater basins. Specifically, basins 1, 3, and 4 (Figure 2) are selected for control because they experience the largest average volume during a storm event. It is assumed that at any time step the agent has knowledge of the water levels and valve positions for each of these basins, as well as the basin between them (basin 2 in Figure 2), thus quadrupling the number of observed states compared to the control of a single basin. The action space must also be reduced to make the problem computationally tractable. For the control of the single basin there are 101 possible actions at any given time step (valve opening with 1% granularity). For 3 controlled basins this increases to 101^3 possible control actions at any given time step. This is not only intractable given our own computational resources, but is well beyond the size of any action space covered in other RL literature. Here, to reduce the action space the agent is allowed to only throttle the valves. Specifically, at any time step the agent can only open or close the valve in 5% increments or leave its position unchanged. This results in only three possible actions for each site and thus 27 (or 3^3) possible actions for the entire network.

The agent receives an individual reward for controlling each basin. These rewards are weighted equally and added together to provide a total reward for controlling the larger system. The reward for controlling each basin is given by:

$$r_4(s_t) = \begin{cases} -c_1 h_t + c_4, & h_t \leq H, f_t \leq F \\ -c_2 h_t^2 + c_3 + c_4, & h_t > H, f_t \leq F \\ -c_1 h_t + (F - f_t)c_5, & h_t \leq H, f_t > F \\ -c_2 h_t^2 + c_3 + (F - f_t)c_5, & h_t > H, f_t > F \end{cases} \tag{12}$$

where reward parameters c=$\{c_1, c_2, c_3, c_4, c_5\}$ are chosen as $\{0.5, 1, 3, 1, 10\}$ to retain the relative scale of the single-basin reward formulations. This reward seeks to accomplish practically identical objectives of the third reward function used in the single-basin control scenario. The difference is the quadratic penalty term that is applied to the height constraint. This modification is made to provide the agent with a more explicit guidance in response to the relatively larger state space compared to the single-basin control scenario. In the rare instance that flooding should occur at one of the basins, agent also receives an additional penalty of -10.

3.8. ***Simulation, Implementation, and Evaluation***

Beyond the formulation of the reward function, the use of RL for the control of stormwater systems faces a number of non-trivial implementational challenges. The first relates to the hydrologic and hydraulic simulation framework, which needs to support the integration of a simulation engine that is compatible with modern RL toolchains. The second challenge relates to the implementation of the actual RL toolchain, which must include the deep neural network training algorithms.

Most popular stormwater modeling packages, such as the Stormwater Management Model (SWMM) Rossman (2010) and MIKE Urban Elliott and Trowsdale (2007) are designed for event based or long-term simulation. Namely, the model is initialized, inputs are selected, and the model run continues until the rainfall terminates or simulation times out. While these packages support some rudimentary controls, the control logic is pre-configured and limited to simple site-scale action, such as opening a valve when level exceed a certain value. The ability to support system-level control logic is limited, let alone the ability to interface with external control algorithms, such as the one proposed in this paper. To that end, we implement a step-wise co-simulation approach that was described in one of our prior studies Mullapudi et al. (2017).

Our co-simulation framework separates the hydraulic solver from the control logic by halting the hydraulic model at every time step. The states from the model (water levels, flows, etc.) are then transferred to the external control algorithm, which makes recommendation on which actions to take (valves settings, pump speeds, etc.). Here, we adopt a python based SWMM package for simulating the stormwater network Riaño-Briceño et al. (2016). This allows the entire toolchain to be implemented using a high-level programming environment, without requiring any major modifications to hydraulic solvers that are often implemented in low-level programming languages and difficult to fuse with modern libraries and open source packages. While other or more complex stormwater or hydrologic models could be substituted, model choice is not necessarily the main contribution of this paper. Rather, we content that SWMM adequately captures runoff and flow dynamics for the purposes of this paper. SWMM models the flow of water in the network using an implicit dynamic wave equation solverRossman (2010). This allows it to effectively model the nuanced conditions (e.g. back channel flows, flooding) that might develop in the network though real time control. Furthermore, the authors have access to calibrated version of the model for this particular study area, which has been documented in a prior study CDMSmith (2015); Wong and Kerkez (2018).

One major task is the implementation of the deep neural network that is used to approximate the RL agent's action value function. Deep neural networks are computationally expensive to train LeCun et al. (2015b). Efficient implementation address this by leveraging a computer's graphical processing unit (GPU) to carry out this training, which is a non-trivial task. To that end, a number of open source and community libraries have emerged, the most popular of which is *TensorFlow* Abadi et al. (2016). This state-of-the-art library has been used in some of the most well-cited RL papers and benchmark problems, which is the reason we choose to adopt it for this study. *TensorFlow* is a python library and can be seamlessly interfaced with our python-based stormwater model implementation.

Four agents are trained and evaluated across the two scenarios: three for the control of individual basins, and one agent for the multi-basin control. A deep neural network is designed and implemented to learn the value function of each agent. The network contains 2 layers with 50 neurons per layer. This network is set up with a *ReLu* activation function Goodfellow et al. (2016) in the internal layers and a linear activation function in the final layer. The full parameters used in the study, including those for gradient descent and the DQN, are provided in appendix B of this paper. A Root Mean Square Propagation (RMSprop) Goodfellow et al. (2016) form of stochastic gradient descent is used for updating the neural network as this variant of gradient descent has been observed to improve convergence.

One storm event is used to train these agents. The SWMM model is forced with a 25-year storm event of 6 $hour$ duration and 63.5 mm intensity (Figure 3). This event generates a total runoff of $3670.639m^3$ with a peak flow of $0.35m^3/s$ at the outlet of watershed. The agents are provided with an operational water level goal H of $2m$, flooding level H_{max} of 3.5m and outflow exceedance threshold of F of $0.10m^3s^{-1}$. It is important to note that the outflow threshold, in particular, serves more as an approximate guide rather than exact requirement, since the discrete valve settings used by the RL agents may not allow the exact setpoint to be physically realizable (e.g. throttling a valve by 5% will limit outflow precision correspondingly). These setpoints are chosen to reflect realistic flooding and stream erosion goals in the study watershed. Agents are trained on a Tesla K20 GPUs on University of Michigan's advanced research computing's high performance cluster.

A fifth agent is also trained, focusing specifically on the multi-basin control scenario. The agent uses the same neural network architecture of the other multi-basin RL control agent, trained this time, however, using *batch normalization* Ioffe and Szegedy (2015). Batch normalization is the process of normalizing the signals between the internal layers of the neural network to minimize the internal covariance shift and has been observed to improve the performance of the deep RL agents Lillicrap et al. (2015). Ioffe et al. (2015) Ioffe and Szegedy (2015) provides a detailed discussion on batch normalization.

The performance of each agent is evaluated by comparing the RL-controlled hydrographs and water levels to those that are specified in the reward functions. For the agents controlling the individual basins this is used to determine the importance of the reward formulation on performance, reward convergence, and training period duration. For the multi-basin control scenario, the same approach is used to quantify overall performance, comparing this time the agent that uses the batch normalized neural network to the agent that uses the non-normalized network.

To evaluate the ability of an RL-agent to control storms that it is not trained on, a final analysis is carried out. Since the agent controlling multiple basins presents the most complex of the scenarios, it is first trained on one of storms and evaluated on a spectrum of storm events with varying return periods (1 to 100 years) and durations (5 min to 24 hours). These storm events are generated based on the SCS type II curve and historical rainfall intensities for the study region Scs (1986). The performance of the agent across these 70 storms is compared to the uncontrolled system to evaluate the boarder benefits of real time control. To allow for a comparison between the controlled and uncontrolled system, a non-negative performance metric is introduced to capture the magnitude and time that the system deviates from desired water level and flows thresholds. Specifically, across a duration T the final performance P adds together the deviation of all N controlled sites from their desired water level (P_h) and flow thresholds (P_f), where:

$$P_h(h) = \begin{cases} h - H, & h > H \\ h - H + 100h, & h > H_{max} \\ 0, & otherwise \end{cases} \tag{13}$$

$$P_f(f) = \begin{cases} 10(f - F), & f > F \\ 0, & otherwise \end{cases} \tag{14}$$

$$P = \sum_{n=1}^{N} \sum_{i=0}^{T} P_h(h_i^n) + P_f(f_i^n) \tag{15}$$

A relatively lower performance value is more desirable, since it implies that the system is not flooding, nor exceeding desired flow thresholds.

4. Results

4.1. Scenario 1: Control of single basin

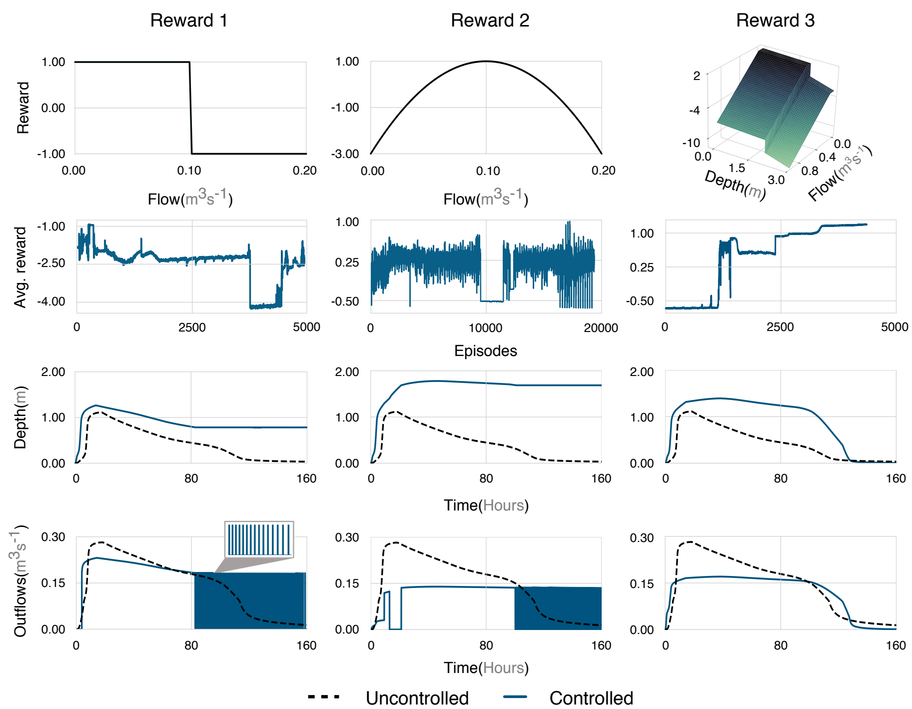

Figure 3: RL control of a single basin, trained using three reward formations (grouped by column). The first row plots each reward function used during training. The second row plots the average reward received during training. The third and fourth rows compare the uncontrolled flows and water levels, for the episode that resulted in the highest reward. Generally, more complex reward formulations lead to relatively better control performance and improved convergence during raining. Agents trained using relatively simple reward also exhibit a rapidly-changing and unstable control behavior, shown as a close up in the bottom left plot.

The ability of an RL agent to control a stormwater basin is highly sensitive to the reward function formulation. Generally, a more complex reward function – one that embeds more information and explicit guidance – performs better, as illustrated in Figure 3. Each column of the figure corresponds with an individual RL agent, each of which is trained using a different reward function (r_1, r_2, r_3).

The reward functions are plotted in the first row, while the reward received during training is plotted in the second row. The training period is quantified in terms of episodes, each of which corresponds to one full SWMM simulation across an entire storm. The third and fourth rows in the figure compare the uncontrolled flows and water levels, respectively, for the episode that resulted in the highest reward.

The RL agent that uses the simplest reward function has the relatively worst performance (Figure 3, first column). Even after 5000 training episodes (a week of real-world simulation time), the mean reward does not converge to a stable value. Playing back the episode that resulted in the highest reward (Figure 3, rows 3-4, column 1), reveals that the RL agent does retain more water than would have been held in the uncontrolled basin. While this lowers the peak flows relative to the uncontrolled basin, the RL agent is generally not able to keep flows below the desired threshold. More importantly, the RL agent's control actions begin oscillating and become unstable toward the middle of the simulation. In this episode, the agent keeps the water level in the basin relatively constant by opening the valve very briefly to release just a small amount of water. This "chattering" behavior (shown as a close up in the figure) results in an unstable outflow pattern that oscillates in a step-wise fashion between $0m^3/s$and $0.18m^3/s$. For various practical reasons, such rapid control actions are not desirable. Since the RL agent never once receives a positive reward, it may have converged onto an undesirable local minimum during the training. Providing more time training does not appear to resolve this issue, which may also suggest that a stable solution cannot be derived using this particular reward formulation.

Increasing the complexity of the reward formulation improves the control performance of the RL agent (Figure 3, second column). When the second and more continuous reward function is used by the agent, the highest reward episode reveals that the RL agent is relatively more effective at maintaining flows at a constant value. Unlike the RL agent using the simple step-wise reward function, the RL agent using the parabolic reward function has more opportunities to receive smaller, more gradual rewards. During most of the episode, this increased flexibility allows the second RL agent to receive positive rewards and keep outflow below a flow threshold of $0.14m^3/s$. While relatively improved, the RL agent using the parabolic reward also does not converge to a stable reward value, however. Toward the end of the episode, this RL agent also carries out irregular and sudden control actions by opening and closing the valve in short bursts. In this case, the RL agent is oscillating between a maximum (valve open) and minimum (valve closed) reward rather than taking advantage of variable rewards in other configurations. This suggests that the agent has either not yet learned a better strategy or, again, that a stable solution cannot be converged upon using this particular reward formulation.

The RL agent using the third and most complex reward function exhibits the relatively best control performance. This agent regulates flow and water levels in a relatively gradual and smooth manner. Unlike in the case of the other two RL agents, after 3500 training episodes the third agent does converge to a steady reward. Evaluating the episode resulting in the highest reward (Figure 3, rows 3-4, column 3), the desired "flat" outflow hydrograph is achieved. No unstable or oscillatory control actions are evident, as in the case of the other two reward functions. The agent is able to maintain flows below a constant threshold of $0.15m^3/s$. While this is not the exact threshold that was specified ($0.1m^3/s$), it is close considering that achievable threshold is dependent on water levels and the ability to only throttle the valve in 1% increments. As stated in the methods section, matching the exact threshold may not be physically realizable in any given situation due to constraints enforced by discretized throttling. Furthermore, the RL agent must balance the desired outflow against the possibility of flooding, and is thus more likely to release a greater amount of water than is specified by the threshold. Interestingly, this agent does not change its valve configuration at all. Rather, it keeps its valve 54% open the entire time of the simulation, which allows it to meet a mostly constant outflow given the specific inflows. Overall, the general shape of the outflows is improved compared to the uncontrolled scenario. Furthermore, an added benefit of real time control is that the overall volume of water leaving the basin is also reduced by 50% due to infiltration.

This scenario, which focuses on the control of a single site, emphasizes the importance of the reward function formulation in RL control of stormwater systems. The complexity of the reward formulation plays an important role in allowing the RL agent to learn a control policy to meet the desired hydrologic outcomes. The importance of reward formulations has been acknowledged in prior studiesSutton and Barto (1998); Ng et al. (1999). Generally, more complex reward function lead to a more rapid convergence of a control policy, while avoiding unintended control actions, such as the chattering behavior seen in figure 3. In fact, prior studies have attributed such erratic control actions to the use of oversimplified reward functions Ng et al. (1999), but have offered little specificity or concrete design recommendations that could be used to avoid such actions. As such, our approach heuristically evaluates reward formulations of increasing complexity until arriving at one that mostly meets desired outcomes. This introduces an element of design into the use of RL for the real time control of stormwater, as the one cannot simply rely on the implicit black box nature of neural networks to solve a control problem under complex system dynamics.

The reward function needs to embed enough information to help guide the RL agent to a stable solution. This introduces only a limited amount of overhead, as reward functions can be intuitively formulated by someone with knowledge of basic hydrology.

For control individual basins, the reward function presented here should be directly transferable. If more complex outcomes are desired, modifications to the reward function may need to be carried out. Objectively, the convergence of the reward will serve as one quality measure of control performance. The ultimate performance of RL for the control of individual sites will, however, need to be assessed on a case-by-case basis by a designer familiar with the application. Taking the baseline lessons learned during the control of a single basin, the second scenario can now evaluate the simultaneous control of multiple basins.

4.2. ***Scenario 2: Control of multiple basins***

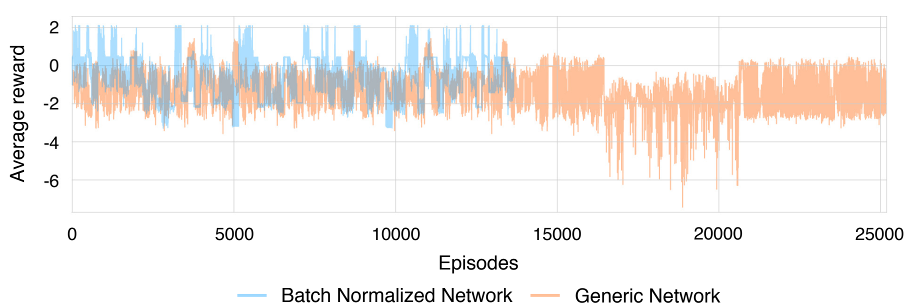

Figure 4: Average reward earned by the RL agent when learning to control multiple basins. The use of neural network batch normalization (blue) leads to consistently higher rewards when compared to the use of a generic neural network (orange). The batch normalized network also leads to higher rewards earlier in the training process.

When trained using the generic feed forward neural network configuration that was used for the control of a single basin, the RL agent controlling multiple assets was unable to converge to a stable reward, even after 25000 episodes of training (Figure 4). This totaled to 52 days of computation time on our GPU cluster, after which the training procedure was halted due to lack of improved reward. Overall, learning performance was low in this configuration. Not only did the learning procedure not converge to stable reward, but the vast majority of rewards were negative. Given this observation, this ineffective neural network was then replaced with one that was batch normalized. The agent using the batch normalized neural network achieved a higher average reward than the agent with a generic feed forward neural network (Figure 4). Furthermore, the agent using the batch normalized neural network achieved a relatively high rewards early on in the training process, thus making it more computationally favorable.

Even with batch normalization, the RL agent did not consistently return to the same reward or improve its performance when perturbed. The exploration in its policy caused the RL agent to oscillate between local reward maxima. Similar outcomes have been observed in a number of RL benchmark problemsHenderson et al. (2017); Mnih et al. (2015), which exhibited a high degree of sensitivity to their exploration policy. Prior studies have noted that the exploration-exploitation balance is difficult to parameterize because neural networks tend to latch onto a local optimum Larochelle et al. (2009). As such, it is likely that the lack of convergence observed in this scenario was caused by the use of a neural network as a function approximator. Forcing neural networks to escape local minima is still an ongoing problem of research Osband et al. (2016). Nonetheless, even without a consistent optimum, the maximum reward obtained during this scenario can still be used as part of an effective control approach.

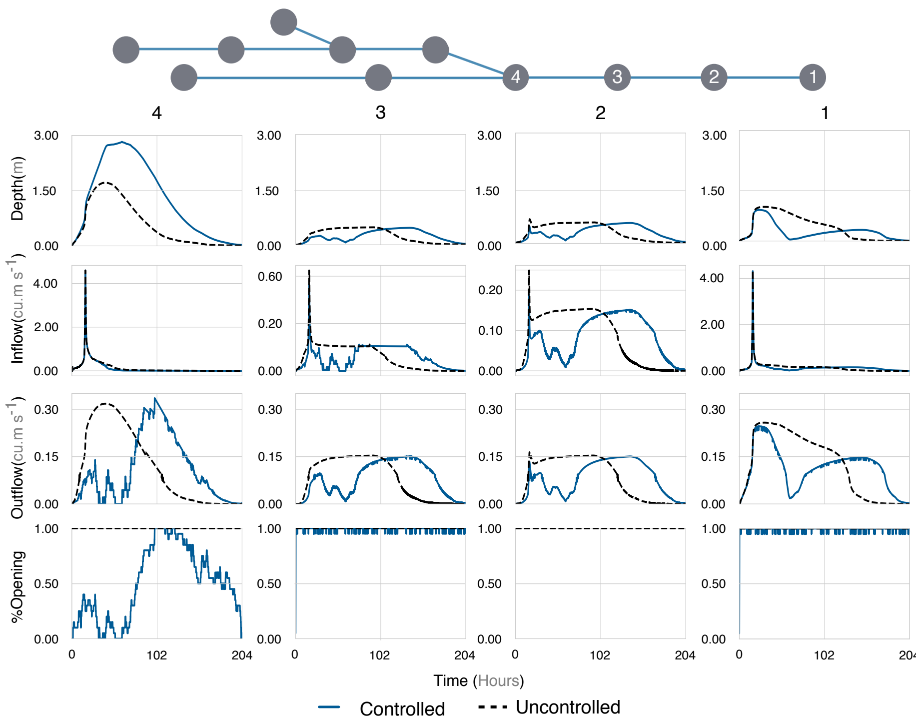

Figure 5: RL agent controlling multiple stormwater basins during a 6-hour, 25-year storm event. Control actions at each of the controlled basins are shown as valve settings in the fourth row of the plot. In this scenario, the agent achieves a high reward by just controlling the most upstream control asset (4) and shifting the peak of the hydrograph.

Selecting the episode with the highest reward revealed the actions taken by the RL agent during the training storm (Figure 5). The figure compares the controlled and uncontrolled states of the four basins during a 25-year 6-hour storm event, showing the depth in each basin, inflows, outflows, and control actions taken by the RL agent. No flooding occurred during this simulation, which means that the reward received by the RL agent was entirely obtained by meeting outflow objectives. The valves on basins 1 and 3 throttled between 100% and 95% open, which for all practical considerations could be considered uncontrolled. As such, the RL agent in this scenario earned its reward by only controlling the most upstream basin in this network.

While the outcome of control was somewhat favorable compared to the uncontrolled systems, the playback of the highest reward in figure 5 does not show drastically different outcomes. Control of the 4^{th} basin shifted the timing of the outflows from the basin but did not reduce its outflows. This resulted in improvements at the 1^{st}, 2^{nd} and 3^{rd} basins. By delaying flows from the 4^{th} basin, the RL agent allowed the downstream basins to drain first and to spend less time exceeding the flow threshold. Interestingly, the RL agent did not control basin 1, even while the single-basin control scenario makes it is clear that a more favorable outcome can be achieved with control (Figure 3). As such, a better control solution may exist, but converging to such a solution using a neural network approximator is difficult. This likely has to do with the larger state action space. While the site-scale RL agent was only observing water level at one basin, the system level RL agent had to track levels and flows across more basins, which increases the complexity of the learning problem. The rewards received by the RL agent in the scenario are cumulative, which means that improvement at just a few sites can lead to better rewards, without the need to control all of them. Increasing the opportunity to obtain rewards thus increases the occurrence of local minima during the learning phase.

In the single basin control scenario the RL agent can immediately observe the impact of its control actions. In the system scale scenario more time is needed to observe water flows through the broader system, which means that the impact of a control action may not be observed until later timesteps. This introduces a challenge, as the RL agent has to learn the temporal dynamics of the system. This challenge has been observed in other RL studies, which have shown better performance for reactive RL problems, as opposed to those that are based on the need to plan for future outcomes Aytar et al. (2018). The need to include planning is still an active area of RL research. Potential emerging solutions include adversarial playSilver et al. (2017b,a), model-based RLClavera et al. (2018), and policy-based learning Schulman et al. (2017). The benefits of these approaches have recently been demonstrated for other application domains and should be considered in the future for the control of water systems.

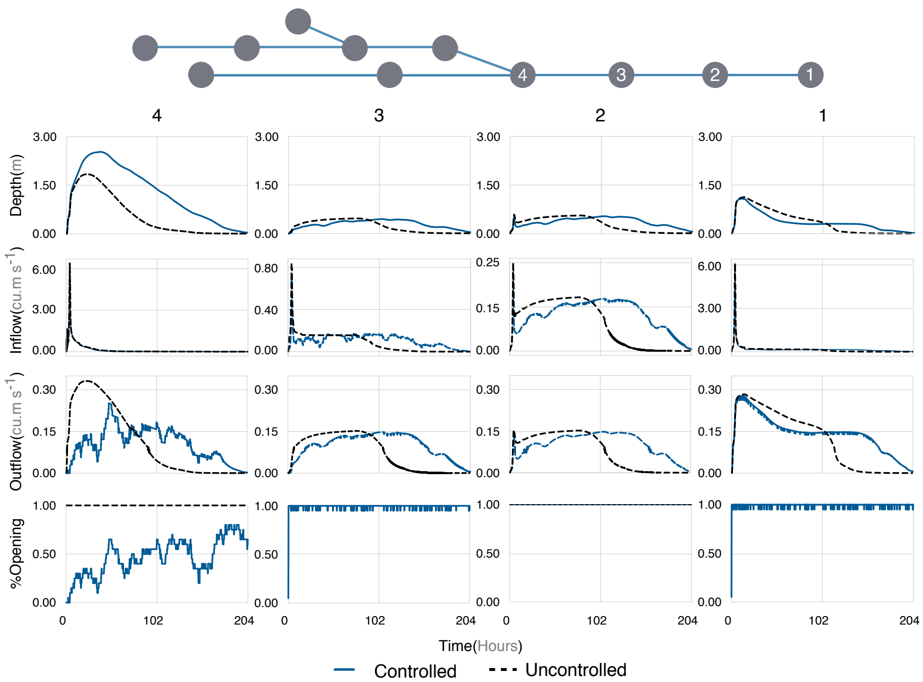

Figure 6: RL agent controlling multiple stormwater basins during a 24-hour, 10-year storm event. Control actions at each of the controlled basins are shown as valve settings in the fourth row of the plot. In this scenario, the agent achieves a high reward, by maximizing the storage utilization in the most upstream control asset (4) and regulating the outflow from it to meet the downstream objectives.

It is important to note that figure 5 represents an evaluation of the RL agent for one storm only – namely, the training storm. Realistically, the control system will need to respond to storms of varying durations and magnitudes. As an example, the RL agent's response to a 24-hour, 10-year storm is shown in figure 6. Here, the RL agent outperformed the uncontrolled system much more notably compared to the training storm. The controlled outflows were much closer to the desired threshold, even when only one basin was controlled. This broader performance is captured in figure 7, which quantifies performance (eq 15) across a spectrum of storm inputs. Figure 7 compares the uncontrolled system to the RL-controlled system. Both the controlled and uncontrolled systems perform equally well during small-magnitude and short events (e.g. the training storm in Figure 5). The benefits of control become more pronounced for larger events, starting at 10-year storms and those that last over 2 hours. This visualization holistically captures the benefits of real time control by highlighting new regions of performance and showing how control can push existing infrastructure to perform beyond its original design.

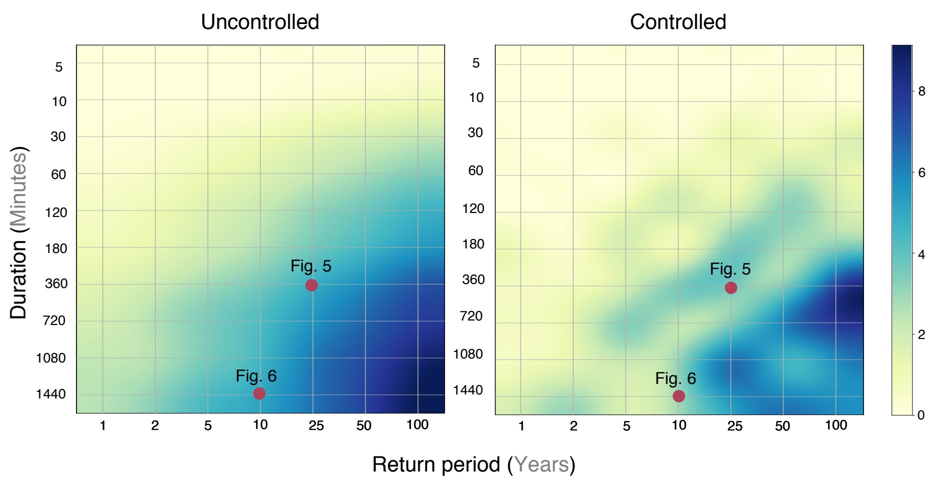

Figure 7: Performance of stormwater system (eq 15) for the uncontrolled system (left) and RL-controlled system (right). The use of RL control enhances the performance stormwater network by allowing the system to achieve desired outcomes across larger and longer storms. The lighter color corresponds with a relatively better performance.

5. Discussion

Given the recent emergence and popularity of Reinforcement Learning, much research still remains to be conducted to evaluate its potential to serve as a viable methodology for controlling water systems. Our study brings to light a number of benefits and challenges associated with this task. Arguably, it seems that the major benefit of using RL to control water systems is the ability to simply hand the learning problem to a computer without needing to worry about the many complexities, nonlinearities and formulations that often complicate other control approaches. However, as this study showed, this comes with a number of considerable caveats. These include the challenges associated with formulating rewards, choosing function approximators, deciding on the complexity of the control problem, as well as contending with practical implementation details.

Our study confirms that the performance of RL-based stormwater control is sensitive to the formulation of the reward function, which has also been observed in other application domains Ng et al. (1999). The formulation of the reward function requires domain expertise and an element of subjectivity, since the RL agent has to be given guidance on what constitutes appropriate actions. In the first scenario it was shown that a reward function that is too simple may lead to adverse behavior, such as the chattering or sudden actions. The reward may also not converge to a stable solution since the neural network can take advantage of the simple objective to maximize rewards using sudden or unintuitive actions. The formulation of the problem, which depends heavily on neural networks, also makes it difficult to determine why one specific reward function may work better than another. Increasing the complexity of the reward function was shown here to help guide the RL agent to a more desirable outcome. Reward formulations are an ongoing research area in the RL community and some formal methods have recently been proposed to provide a more rigorous framework for reward synthesis Fu et al. (2017). These formulations should be investigated in the future.

Even when the choice of reward function is appropriate or justifiable, the control performance can become sensitive to the approximation function, which in our case took the form of a deep neural network. Choosing the architecture and structure of the underlying network becomes an application dependent task and can often only be derived through trial and error Sutton and Barto (1998); Henderson et al. (2017). Secondly, for challenging control problems, such as the one studied here, learning the mapping between rewards and all possible control decisions becomes a complex task. The neural network must be exposed to as many inputs and outputs as possible, which is computationally demanding. In our study we ran simulations for many real-world months on a high performance cluster, but it appears that the learning phase could have continued even longer. This, in fact, has been the approach of many successful studies in the RL community, where the number of computers and graphical processing units can be in the hundreds Espeholt et al. (2018); OpenAI (2018). This was not feasible given our own resources, but could be evaluated in the future.

Aside from the formulation of the learning functions and framework, the actual complexity and objectives of the control problem may pose a barrier to implementation. We showed that an RL agent can learn how to control a single stormwater basin effectively, but that controlling many sites at the same time is difficult. A major reason is the increase in the number of states and actions that must be represented using the neural network. While computational time may remedy this concern, the structure of the neural network may also need to be altered. In a system-scale stormwater scenario, actions at one location may influence another location at a later time. As such, the agent would benefit from a planning-based approach which considered not only current states, but future forecasts as well. Such planning-based approaches have been proposed in the RL literature and should be investigated to determine if they lead to an improvement in performance Clavera et al. (2018); Depeweg et al. (2016).

Furthermore, model-based approaches have also recently been introduced and could allow some elements neural network to be replaced with an actual physical or numerical water model Gu et al. (2016). Such approaches should be evaluated in the future since they may permit more domain knowledge from water resources to be embedded in the learning problem.

Finally, the use of RL for the control of stormwater systems is underpinned by a number of practical challenges. Computational demands are very high, especially compared to competing approaches, such as dynamical systems control, model predictive control, or market-based controls. While computational resources are becoming cheaper, the resources require to carry out this study were quite significant and time demanding. Since actions taken by neural networks cannot easily be explained and explicit guarantees cannot be provided, this may limit adoption by decision makers who may consider the approach a "black box". It is also unlikely that the control of real-world stormwater systems will simply be handed over to a computer that learns through mistakes. Rather, simulation-based scenarios will be required first. It has recently been shown as long as a realistic simulator is used - in our case SWMM - then the agent can be effectively trained in a virtual environment before refining its strategy in the real world OpenAI (2018).

6. Conclusion

This paper introduced an algorithm for the real time control of urban drainage systems based on Reinforcement Learning (RL). While RL has been used successfully in the computer science communities, to our knowledge this is the first instance for which it has been explicitly adopted for the real time control of urban water systems. The methodology and our implementation show promise for using RL as an automated tool-chain to learn control rules for simple storage assets, such as individual storage basin. However, the use of RL for more complex system topologies faces a number of challenges, as laid out in the discussion. Simultaneously controlling multiple distributed stormwater assets across large urban areas is a non-trivial problem, regardless of the control methodology. To that end, the concepts, initial results and formulations provided by this paper should help build a foundation to support RL as a viable option for stormwater control. The source code accompanying this paper should also allow others to evaluate many other possible architectures and parameterizations that could be used to improve the results presented in the paper.

Acknowledgements

This research was supported by Great Lakes Protection Fund (grant number # 1035) and the US National Science Foundation (grant number # 1737432). We also would like to acknowledge the support provided by Advanced Research Computing at the University of Michigan, Ann Arbor.

Appendix A: Deep Neural Networks

Broadly, neurons are the fundamental processing elements of a neural network. They receive their inputs (x_{ij}) as the weighted (w_{ij}) outputs from the neurons in the previous layers and produce a single output signal (y_j) dependent upon a bias (b_j) and activation function $f(*)$

$$z_j = \sum_{i}^{n} w_{ij} x_{ij} + b_j \tag{A1}$$

$$y_j = f(z_j) \tag{A2}$$

More specifically, "deep" neural networks are a collection of such neurons organized as distinct layers. A neural network approximates a function by fine tuning its weights and biases so that its output signal closely resembles the output of the function it is approximating for the same inputs. The degree of resemblance between the signals is computed based on a loss function $L(*)$

$$Loss = L(Q_p, Q_o) \tag{A3}$$

$$w_{ij} = w_{ij} - \alpha \times \frac{dL(Q_p, Q_o)}{(dw_{ij})} \tag{A4}$$

$$b_j = b_j - \alpha \times \frac{dL(Q_p, Q_o)}{(db_j)} \tag{A5}$$

The choice of the loss function is dictated by the nature of the function being approximated. For example, a neural network approximating rainfall runoff may use mean squared error Tokar and Johnson (1999)

$$Loss = |Q_p - Q_o|^2 \tag{A6}$$

The closer the correspondence between the signals, lower the loss.

Neural networks minimize the loss though stochastic gradient descent, starting with a set of weights and biases (either random or values sampled from a distribution). Based on the value of loss function, the values of weights and biases are adjusted, and the neural network attempts to approximate the function with these updated values. This process of tuning the weights and biases is repeated until the neural network can approximate the function to satisfaction or loss is minimized. While deep neural networks show significant promise in approximating functions, their ability to do so is contingent upon several factors: Stability of the learning process, the size and depth of the network, the underlying data distributions Goodfellow et al. (2016); Chollet (2017). Fundamental description of deep neural networks can be found in established textbooks Goodfellow et al. (2016).

Appendix B: Hyper parameters and architecture

Neural Network

Layers	2
Number of Neurons per layer	50

Gradient Descent

Learning Rate	10^{-3}
Rho	0.9
Epsilon	10^{-8}
Decay	0.0

Batch Normalization

Momentum	0.99
Epsilon	0.001

Deep Q Network

Target Network Update	10000
Gamma	0.99
Replay Buffer	100000

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al., 2016. Tensorflow: a system for large-scale machine learning., in: OSDI, pp. 265–283.

Abdulhai, B., Kattan, L., 2003. Reinforcement learning: Introduction to theory and potential for transport applications. Canadian Journal of Civil Engineering 30, 981–991.

Aytar, Y., Pfaff, T., Budden, D., Paine, T.L., Wang, Z., de Freitas, N., 2018. Playing hard exploration games by watching youtube. arXiv preprint arXiv:1805.11592 .

Bartos, M., Wong, B., Kerkez, B., 2017. Open storm: a complete framework for sensing and control of urban watersheds. arXiv preprint arXiv:1708.05172 .

Bhattacharya, B., Lobbrecht, A., Solomatine, D., 2003. Neural networks and reinforcement learning in control of water systems. Journal of Water Resources Planning and Management 129, 458–465.

Castelletti, A., Galelli, S., Restelli, M., Soncini-Sessa, R., 2010. Tree-based reinforcement learning for optimal water reservoir operation. Water Resources Research 46.

CDMSmith, 2015. City of ann arbor stormwater model calibration and analysis project. Accessed: 2018-09-25.

Chollet, F., 2017. Deep learning with python. Manning Publications Co.

Clavera, I., Rothfuss, J., Schulman, J., Fujita, Y., Asfour, T., Abbeel, P., 2018. Model-based reinforcement learning via meta-policy optimization. arXiv preprint arXiv:1809.05214 .

Depeweg, S., Hernández-Lobato, J.M., Doshi-Velez, F., Udluft, S., 2016. Learning and policy search in stochastic dynamical systems with bayesian neural networks. arXiv preprint arXiv:1605.07127 .

Elliott, A., Trowsdale, S., 2007. A review of models for low impact urban stormwater drainage. Environmental Modelling & Software 22, 394–405. URL: https://www.sciencedirect.com/science/article/pii/S1364815206000053, doi:10.1016/J.ENVSOFT.2005.12.005.

Emerson, C.H., Welty, C., Traver, R.G., 2005. Watershed-Scale Evaluation of a System of Storm Water Detention Basins. Journal of Hydrologic Engineering 10, 237–242. URL: http://ascelibrary.org/doi/10.1061/%28ASCE%291084-0699%282005%2910%3A3%28237%29, doi:10.1061/(ASCE)1084-0699(2005)10:3(237).

Espeholt, L., Soyer, H., Munos, R., Simonyan, K., Mnih, V., Ward, T., Doron, Y., Firoiu, V., Harley, T., Dunning, I., et al., 2018. Impala: Scalable distributed deep-rl with importance weighted actor-learner architectures. arXiv preprint arXiv:1802.01561 .

Fu, J., Luo, K., Levine, S., 2017. Learning robust rewards with adversarial inverse reinforcement learning. arXiv preprint arXiv:1710.11248 .

Goodfellow, I., Bengio, Y., Courville, A., Bengio, Y., 2016. Deep learning. volume 1. MIT press Cambridge.

Gu, S., Lillicrap, T., Sutskever, I., Levine, S., 2016. Continuous deep q-learning with model-based acceleration, in: International Conference on Machine Learning, pp. 2829–2838.

Henderson, P., Islam, R., Bachman, P., Pineau, J., Precup, D., Meger, D., 2017.

Deep reinforcement learning that matters. arXiv preprint arXiv:1709.06560 .

Hornik, K., Stinchcombe, M., White, H., 1989. Multilayer feedforward networks are universal approximators. Neural networks 2, 359–366.

Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167 .

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. Environmental Science & Technology 50, acs.est.5b05870. URL: http://pubs.acs.org/doi/abs/10.1021/acs.est.5b05870, doi:10.1021/acs.est.5b05870.

Kober, J., Bagnell, J.A., Peters, J., 2013. Reinforcement learning in robotics: A survey. The International Journal of Robotics Research 32, 1238–1274. URL: http://journals.sagepub.com/doi/10.1177/0278364913495721, doi:10.1177/0278364913495721.

Laris Karklis, K.U., Muyskens, J., 2017. Before-and-after photos of Harvey flooding in Texas - Washington Post. URL: https://www.washingtonpost.com/graphics/2017/national/harvey-photos-before-after/?utm_term=.5976b1e7a7ed.

Larochelle, H., Bengio, Y., Louradour, J., Lamblin, P., 2009. Exploring strategies for training deep neural networks. Journal of machine learning research 10, 1–40.

LeCun, Y., Bengio, Y., Hinton, G., 2015a. Deep learning. Nature 521, 436–444. URL: http://www.nature.com/articles/nature14539, doi:10.1038/nature14539.

LeCun, Y., Bengio, Y., Hinton, G., 2015b. Deep learning. nature 521, 436.

Lillicrap, T.P., Hunt, J.J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., Wierstra, D., 2015. Continuous control with deep reinforcement learning. arXiv preprint arXiv:1509.02971 .

Mahmoodian, M., Delmont, O., Schutz, G., 2017. Pollution-based model predictive control of combined sewer networks, considering uncertainty propagation. International Journal of Sustainable Development and Planning 12, 98–111. URL: http://www.witpress.com/journals/SDP-98-111, doi:10.2495/SDP-V12-N1-98-111.

Meneses, E., Gaussens, M., Jakobsen, C., Mikkelsen, P., Grum, M., Vezzaro, L., 2018. Coordinating Rule-Based and System-Wide Model Predictive Control Strategies to Reduce Storage Expansion of Combined Urban Drainage Systems: The Case Study of Lundtofte, Denmark. Water 10, 76. URL: http://www.mdpi.com/2073-4441/10/1/76, doi:10.3390/w10010076.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., Riedmiller, M., 2013. Playing Atari with Deep Reinforcement Learning. arXiv preprint arXiv: . . . , 1–9URL: http://arxiv.org/abs/1312.5602, doi:10.1038/nature14236.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.a., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., Hassabis, D., 2015. Human-level control through deep reinforcement learning. Nature 518, 529–533. URL: http://dx.doi.org/10.1038/nature14236, doi:10.1038/nature14236.

Mollerup, A.L., Mikkelsen, P.S., Thornberg, D., Sin, G., 2016. Controlling sewer systems – a critical review based on systems in three EU cities. http://dx.doi.org/10.1080/1573062X.2016.1148183 .

Mullapudi, A., Wong, B.P., Kerkez, B., 2017. Emerging investigators series: building a theory for smart stormwater systems. Environ. Sci.: Water Res. Technol. URL: http://xlink.rsc.org/?DOI=C6EW00211K, doi:10.1039/C6EW00211K.

Ng, A.Y., Coates, A., Diel, M., Ganapathi, V., Schulte, J., Tse, B., Berger, E., Liang, E., 2006. Autonomous inverted helicopter flight via reinforcement learning, in: Experimental Robotics IX. Springer, pp. 363–372.

Ng, A.Y., Harada, D., Russell, S., 1999. Policy invariance under reward transformations: Theory and application to reward shaping, in: ICML, pp. 278–287.

Ogata, K., 2011. Modern Control Engineering. 5 ed., Prentice Hall. URL: https://books.google.com/books?id=Wu5GpNAelzkC&q=Modern+Control+Engineering&dq=Modern+Control+Engineering&hl=en&sa=X&ved=0ahUKEwjtzbbizbfOAhVU-GMKHZyOChcQ6AEIHDAA.

OpenAI, 2018. Openai five.

Osband, I., Blundell, C., Pritzel, A., Van Roy, B., 2016. Deep exploration via bootstrapped dqn, in: Advances in neural information processing systems, pp. 4026–4034.

Pleau, M., Colas, H., Lavallee, P., Pelletier, G., Bonin, R., 2005. Global optimal real-time control of the quebec urban drainage system. Environmental Modelling & Software 20, 401–413.

Riaño-Briceño, G., Barreiro-Gomez, J., Ramirez-Jaime, A., Quijano, N., Ocampo-Martinez, C., 2016. MatSWMM – An open-source toolbox for designing real-time control of urban drainage systems. Environmental Modelling & Software 83, 143–154. URL: https://www.sciencedirect.com/science/article/pii/S1364815216301451, doi:10.1016/J.ENVSOFT.2016.05.009.

Rossman, L.A., 2010. Storm Water Management Model User's Manual Version 5.1. US Environmental Protection Agency.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O., 2017. Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 .

Schütze, M., Campisano, A., Colas, H., Schilling, W., Vanrolleghem, P.A., 2004. Real time control of urban wastewater systems—where do we stand today? Journal of Hydrology 299, 335–348. doi:10.1016/j.jhydrol.2004.08.010.

Scs, U., 1986. Urban hydrology for small watersheds, technical release no. 55 (tr-55). US Department of Agriculture, US Government Printing Office, Washington, DC .

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., et al., 2017a. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. arXiv preprint arXiv:1712.01815 .

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al., 2017b. Mastering the game of go without human knowledge. Nature 550, 354.

Sutton, R., Barto, A., 1998. Reinforcement learning. MIT Press, Cambridge.

Sutton, R.S., 1991. Planning by incremental dynamic programming, in: Machine Learning Proceedings 1991. Elsevier, pp. 353–357.

Sutton, R.S., McAllester, D.A., Singh, S.P., Mansour, Y., 2000. Policy gradient methods for reinforcement learning with function approximation, in: Advances in neural information processing systems, pp. 1057–1063.

The European Parliament and the council of European Union, 2000. Directive 2000/60/EC of the European Parliament and of the Council of 23 October 2000 establishing a framework for Community action in the field of water policy. Official Journal of the European Communities , 01 – 73URL: http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32000L0060.

Tokar, A.S., Johnson, P.A., 1999. Rainfall-runoff modeling using artificial neural networks. Journal of Hydrologic Engineering 4, 232–239.

Van Overloop, P.J., 2006. Model predictive control on open water systems. IOS Press.

Vezzaro, L., Grum, M., 2014. A generalised dynamic overflow risk assessment (dora) for real time control of urban drainage systems. Journal of Hydrology 515, 292–303.

Watkins, C.J.C.H., Dayan, P., 1992. Q-learning. Machine Learning 8, 279–292. URL: http://link.springer.com/10.1007/BF00992698, doi:10.1007/BF00992698.

Watson, S.B., Miller, C., Arhonditsis, G., Boyer, G.L., Carmichael, W., Charlton, M.N., Confesor, R., Depew, D.C., Höök, T.O., Ludsin, S.A., Matisoff, G., McElmurry, S.P., Murray, M.W., Peter Richards, R., Rao, Y.R., Steffen, M.M., Wilhelm, S.W., 2016. The re-eutrophication of Lake Erie: Harmful algal blooms and hypoxia. Harmful Algae 56, 44–66. URL: https://www-sciencedirect-com.proxy.lib.umich.edu/science/article/pii/S1568988315301141, doi:10.1016/J.HAL.2016.04.010.

Wong, B., Kerkez, B., 2018. Real-time control of urban headwater catchments through linear feedback: performance, analysis and site selection. Water Resources Research .