NOTATION TRENDPLUS PUBLISHING
WHERE IDEAS MEET IMPACT

Volume-1, Issue 1, August 2026, Page No- 36-51

Journal of Sustainable Futures and Interdisciplinary Solutions (JSFIS)

Few-Shot Meta-Learned Safety Shields for Real-Time Adaptive Optimization in Sustainable Smart City Digital Twins

Ankur Jain [1*], Shivaji ji[1]

[1]*Assistant professor, Quantum University, Roorkee, Uttarakhand-247662.*

Article Received: 05-04-2026

Revised Version: 28-06-2026

Accepted: 04-07-2026

*Corresponding Author

Ankur Jain

E-Mail Id: ankurjain.me@quantumeducation.in

Abstract: This paper presents a meta-learned constrained reinforcement learning orchestrator that replaces traditional static mixed-integer linear programming solvers for sustainable urban resource management. The conventional approach to coordinating water, energy, and waste systems across heterogeneous districts suffers from rigid constraint formulations that cannot adapt to rapidly changing environmental conditions and demand patterns. Our core innovation is a task-conditioned constraint module that dynamically reformulates sustainability boundaries—such as minimum water pressure or maximum transformer load—based on real-time telemetry, surrogate model outputs, and spectral entropy of consumption signals. These adaptive constraints are integrated into the reinforcement learning objective through a Lagrangian barrier formulation, which a PID controller dynamically weights. Furthermore, we introduce a differentiable physics-informed neural network that continuously predicts resource state evolution and outputs a violation risk field thirty seconds into the future. This risk field then serves as a corrective shielding signal during a model-agnostic meta-learning outer loop, which fine-tunes the policy parameters every five minutes using only sixty-four recent transitions from the target district. The meta-controller executes asynchronously on cloud GPUs and pushes updated policies to edge nodes via gRPC, thereby enabling sub-second policy adaptation without interrupting real-time control. We demonstrate that this architecture achieves few-shot adaptation across multiple urban districts without retraining from scratch, substantially reducing constraint violations during distribution shifts compared to a fixed-solver baseline.

Keywords: Few-Shot Learning, Meta-Learning, Safety Shields, Digital Twins, Sustainable Smart Cities.

Introduction

The concept of smart city is closely related to the need for coordinating and adaptively managing urban systems that intersect, including those of urban water resources (water distribution systems), power resources (electric network), and waste resources (waste treatment systems) (El-Agamy et al., 2024). One of the emerging technologies which has come to the forefront for the realization of such a vision is Digital Twin technology, defined as the ability to create high fidelity variants of the physical infrastructure which allow for its simulation, monitoring and optimization of resource flows in real-time (El-Agamy et al., 2024). Due to the complexity of the urban area, the various district demand patterns and uncertainty of the disturbances, it is not easy to adopt traditional methods of optimization.

In current control technologies for resource coordination in cities, for example model predictive control (MPC) and mixed-integer linear programming (MILP), fixed safety limits are quantified by static constraints (Liu et al., 2020). These types of barriers prevent, for example, the minimum water pressure from a transformer, and are normally designed with conservative engineering considerations, and are almost never varied, if ever, during operation. That isn't good for them when there is a massive block of energy demand following a heat wave, or a waterfall giving way later in the week, or an increased water demand in the summer compared to winter. This often leads to incorrect resources being allocated to the task or worse still safety problems.

The method to tackle the above problems is the introduction of a meta-learned constrained reinforcement learning (RL) controller, also referred to as a meta-learned constrained reinforcement learning orchestrator, that can adaptively reformulate the sustainability-aware constraints based on real-time telemetry information on the environment. Finally, our novelty is a task-conditioned constraint module, that can be adapted and inferred from Digital Twin surrogate models, spectral entropy of consumption signals or from multi-modal sensor observations. Differentiable physics-informed neural network (PINN) (Choi et al., 2022) is used to generate a violation risk field by predicting up to 30 seconds in advance the evolution of the resource states (tank levels, voltage magnitudes). Next, this risk field is integrated into RL target with a Lagrangian barrier method that operates the trade-off weight with a PID control to adapt to the increased risk.

The second major contribution is doing so by taking an adaptive constraint approach and fusing it with model-agnostic meta-learning (MAML) (Jadhav et al., 2025). We split the entire control cycle in two parts, the outer one is the time spent executing the meta-controller in order to set the parameters of the policy every 5 minutes using a small batch (of 64 transitions) of the most recent batch of data from some target district. That is why it is important to be able to quickly adapt to new contexts by performing a few shots of gradient update, without having to tedious retrain from the scratch. This orchestrator is asynchronously running and deployed on cloud GPUs and informs the edge nodes with less than sub-second latency via gRPC for policies changes.

There are some basic differences between the proposed approach, and the previous ones. First, while some safe RL algorithms work with fixed constraints (as in Yang et al., 2026), our approach uses learnable and adaptive constraints that are based on the real-time sensor data streams and vary as needed. Second, we not only adapt the policy (as in typical meta-learning use cases) but also the constraint formulation. Since the PINN based risk predictor is capable of giving feedback on the gradients, it is also possible to adjust the constraint thresholds in the outer loop to either widen or narrow the thresholds in a gradient-driven manner to provide closed-loop control.

Section 2 reviews related research done on DTs in smart cities, constricted RL and meta-learning. Some background on Markov decision process with constraints and meta-learning is provided in Section 3.

In Section 4, we describe the proposed architecture, which has a task-conditioned constraint module, a PINN predictor in an outer loop and a meta-learning loop. In section 5, the experimental set-up and the urban resource simulation in a multi-district system are described. The results on the adaptation speed, constraint satisfying and computational efficiency are reported in Section 6. The implications, limitations and future directions of research are presented in section 7. The paper is concluded with Section 8.

Related Work

Constrained Reinforcement Learning for Safety-Critical Control

There is an increased need for introducing the operating constraints directly in the respective methods, when they are applied to a safety critical system such as the power grid, the water grid, etc. Constrained policy optimization (CPO) (Yang et al., 2026), is on the other hand principled, meaning that constraints can be projected on a trust region and that the a priori boundary of this region can guarantee constraint satisfaction. Lagrangian methods are also adaptive, and allow for punishment of constraint violation through weights but are fixed (Achiam & Amodei, 2019). These approaches have been formulated and implemented in the microgrid control domain with safety shields to avoid actions that otherwise may cause the system voltage to go outside the limits (Nekoei et al., 2025). Such shields are usually handcrafted or formal verified rules, however, and thus do not adapt to the variability in environmental conditions or different dynamics of the various systems.

Digital Twins and Urban Resource Optimization

Digital Twins can offer a real-time virtual model of any real-world system, allowing the urbansphere resources to be modelled, monitored, and predicted in real time (El-Agamy et al., 2024). Then, in the future, the RL agent has been tried to be adapted for traffic management (Salunke, 2023), delivering energy services in energy systems (Rajaram & Almalki, 2026), and Coordinated Water-Energy-Waste systems (El-Agamy et al., 2024). Standards include specifying fixed optimization solvers with a fixed objective function and/or fixed constraints and not being able to modify it during the optimization process in the code. Some frameworks provide some sort of adaptability via changing parameters, or retraining the model, but do not make the boundaries on constraints learnable parameters that could adapt from district to district and to environmental conditions. In addition, the literature does not focus on the different districts with few data which do not require retraining.

Meta-Learning for Fast Adaptation in Control

Model-agnostic meta-learning (MAML) (Jadhav et al., 2025) has shown to work well in control problems that need fast adaptation to the new dynamics (or reward space). For example, in robotics, one has used MAML to take a policy learnt on a terrain to another one using only a few gradient steps (Andrychowicz et al., 2016). Meta-learning has proposed to personalise demand response policy of the new buildings with small amount of data for smart grid applications (Gao et al., 2026). But such approaches often make the assumption that constraints of safety are the same for all tasks. Compared to this, quite little work has been done yet on meta learning (combining the two used here), and adapting the boundary of constraints (as did in many methods) and the policy. Second, existing meta-learning approaches do not leverage predictions for future violation of constraints for adaptation which is inspired by physics.

Physics-Informed Neural Networks for Predictive Modeling

PINNs can obtain internal fields or fluxes accurately from a few measurements by directly adding physics (especially PDEs) into the loss function of the NN (Choi et al., 2022). They are used in the field of hydraulic networks (Bella et al., 2026) and power networks (Falas et al., 2026). In most PINN applications the focus isn't on realtime control, but rather rather on an offline simulation or state estimation. Both, we argue - and provide PINns with an online connection as a risk predictor, too (apart from an inner loop), thus, with differentiable Violation Risk fields, and use them in the meta-learning outer loop for constraint adaptation.

Comparison with Existing Works

The current work's contribution is to overcome the gap between the above works, that is, adaptation of the control policy in the few-shot setting and the safety constraint formulation. In contrast to fixed constraint in CPO (Yang et al., 2026) or other safe RL approaches, we train our spectrum-conditioned constraint module with adaptive thresholds that are dependent on the spectrum entropy and environment. Our outer loop also learns parameters of the constraint generator to learn the district-specific safety boundaries, and MAML adapts policies for new tasks. While PINNs work towards predicting the physical values, our model will predict the occurrence of risk of violation, which will have a direct impact on the shape of the RL objective. Generating a shield is not the same as shielded CU (Nekoei et al., 2025) which is pre-computable, nor that of an offline process. A key novelty of our approach is this combination of meta-learned policy and the adaptive constraints dynamically learned from the PINNs, which make our approach easy to deploy on a large number of different and heterogeneous urban districts without the need of manually re-tuning policy-rules.

Preliminaries

In this section, the background and basic concepts are provided, to aid the understanding of the proposed meta-learned adaptive constraint framework. We firstly formalize the sequential decision making problem with safety concern, physics-informed state prediction modelling, and then the gradient-based paradigm for fast adaptation.

Foundations of Constrained Markov Decision Processes

The characteristic which naturally models sequential decision making tasks with safety requirements corresponds to urban resource management tasks such as coordination of different water pressure, management of the energy load or the waste collection. This type of problems can be typically represented in the form of the Constrained Markov Decision Process (CMDP) (Gladin et al., 2023). The CMDP can be described as a 5-tuple (., ., ., .,), where the state space are, the action space are, the transition probability are, the reward function is a function specifying the reward for taking an action in a state and the cost function is the cost function giving the cost of violation of a constraint; the discount factor and constraint violation threshold . The agent's job is to find a policy that will maximize the EDR with a constraint on the EDC :

$$\max_{\pi} \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^{t}\, R(s_t, a_t)\right] \quad \text{subject to} \quad \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^{t}\, C(s_t, a_t)\right] \leq d \qquad (1)$$

The unconstrained way to solving a CMDP is to convert the constrained minimization problem into an unconstrained one with the use of Lagrangian relaxation. The Lagrangian function is defined as following:

$$\mathcal{L}(\pi, \lambda) = \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t \left(R(s_t, a_t) - \lambda(C(s_t, a_t) - d)\right)\right] \quad (2)$$

where the number is the lagrange multiplier. If it exists, the saddle-point solution is then termed as the optimal policy. In traditional safe RL algorithms (Yang et al., 2026), the constraint function is specified beforehand, and it represents static limits on the operation of the system like the maximum allowable pressure in a pipe or the minimum level in a reservoir. This is a problem in smart city environment as these borders too can be the subject of telemetry and future prediction.

Physics-Informed Machine Learning and Differential Equations

The accuracy of the prediction of evolution of the state of the resources is crucial for pro-active constraint enforcement. Many physical phenomena related to urban infrastructure can be modelled by an equation of the form: , where x is a vector of variables, u(x, t) represents the state of the phenomenon over time, and D[u](x, t) is a function of u, x and t:

$$\frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} = s(u) \quad (3)$$

In this case the state of the resource (such as pressure, voltage, volume of waste) at point is represented by and is a flux function that represents transport, and is a sink or source due to consumption or generation. To achieve this objective, these physical laws are hard coded in this learning objective via Physics-Informed Neural Networks (PINNs) (Choi et al., 2022). The function is parameterized by and a Neural Network (NN) is trained to meet the given PDE and set of experimental observations. The total loss function is the sum of a data-fidelity term and a physics residual term:

$$\mathcal{L} = \frac{1}{N_d}\sum_{i=1}^{N_d}\left|u_{NN}(x_i, t_i) - u_{obs,i}\right|^2 + \frac{1}{N_r}\sum_{j=1}^{N_r}\left|\mathcal{R}\left(u_{NN}(x_j, t_j)\right)\right|^2 \quad (4)$$

Here, - - denotes the number of data points in the observation equation, - - the number of collocation points for the physics residual and is the residual of the PDE in Equation 3. The ability of PINNs to produce accurate predictions with just a few data points and to satisfy known physics makes them perfect for the Digital Twin application where full state observations might not be possible (Bella et al., 2026).

Principles of Gradient-Based Meta-Learning

It's difficult to do a localization and delivery of a policy in real time in a larger smart city and typically, a policy requires many time retraining. In response to this, gradient-based meta-learning was proposed, which approached the problem by pre-training a policy parameterisation that can then be fine-tuned with just a few gradient steps to a new task, which is known as Model-Agnostic Meta-Learning (MAML) (Jadhav et al., 2025). Given a distribution of duties and corresponding loss functions from the observations to the solution sets, MAML sets out to find the initial parameters which can then be fine-tuned into task-specific parameters for each duty to achieve:

$$\min_{\theta} \mathbb{E}_{\mathcal{T}_i \sim p(\mathcal{T})}\left[\mathcal{L}_{\mathcal{T}_i}\left(f_{\theta'_i}\right)\right] \quad \text{subject to} \quad \theta'_i = \theta - \alpha \nabla_{\theta} \mathcal{L}_{\mathcal{T}_i}(f_{\theta}) \quad (5)$$

Here, the learning rate of the inner loop is denoted as and the parameters to learn for task are denoted as . By optimising over the meta-loss which is summed over tasks, the outer-loop optimization (over) runs a search of an initial point in the parameter space to help quickly fine tune the performance for each task. Typically, standard meta-RL uses will only use the expected return in the loss function

(Andrychowicz et al., 2016). However to enable safe control of the target system in the urban setting adaptable constraint descriptions are required, and they must also be differentiable and learnable in the meta-Learning loop. Another major issue whose solution is resolved using the proposed architecture.

Meta-Learning with Adaptive Sustainability Constraints

The framework of the proposed Optimization Orchestrator which consists of a task-conditioned constraint module, differentiable physics-informed neural network (PINN) for risk prediction, and a few-shot meta-learning outer loop for adaptation across heterogeneous urban districts is detailed in this section. The mentioned system is based on a federated clouds-edge Digital Twin architecture (as depicted in the Figure 1).

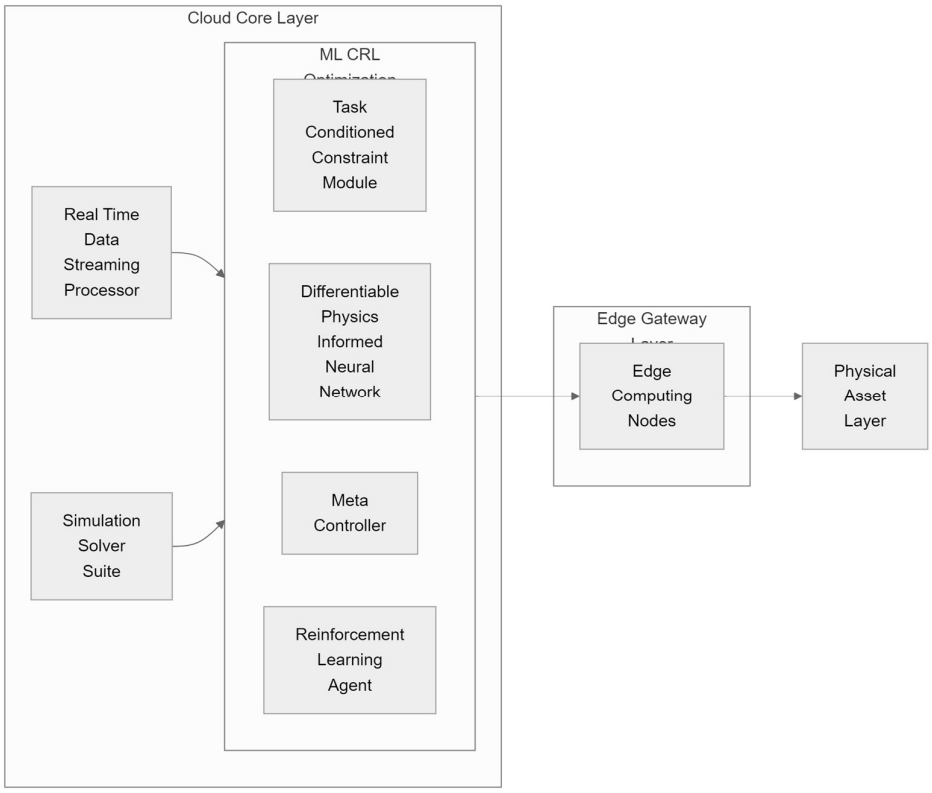

Figure 1. System Integration of ML CRL Optimization Orchestrator

The proposed Orchestrator is going to be put in place in the Digital Twin pipeline, which means that it will be receiving real time data streams from the sensor layer, and the surrogate models from the simulation suite, and performing results transmission to the edge gateway layer. Here it is the "10,000" is actually the definition of an asynchronous such a meta-learning loop, where the training of the orchestrator is done on cloud GPUs while the training of the inner-loop policy is carried out on edge with sub-second latency.

Task-Conditioned Constraint Generation

What is different from the standard constrained RL is the way that the boundaries of the constraint are formulated. We discard the static lower-bound $ \underline{b} $ and upper-bound $ \overline{b} $, and only keep a learned lower-bound generator network $ l(\cdot) $ and upper-bound generator network $ u(\cdot) $ parameterised by . This network also generates dynamic bounds, dynamically adapting to a

vector that represents the current environmental situation, and includes also the current demand patterns, to be provided as a task descriptor.

The task descriptor has three parts; the raw observation from the sensor, a Digital Twin surrogate state and a spectral entropy feature which aggregates the fluctuations of the previous consumption measurement signals, all are updated at each time step. Then, the spectral entropy is:

$$\mathrm{SE}(\mathbf{x}_t, \mathbf{s}_t) = -\sum_f p\,(f)\log p(f), \quad p(f) = \frac{|X(f)|^2}{\sum_{f'} |X(f')|^2} \qquad (6)$$

Fourier Transform of a sliding window frame of the combined signal (sensor signal and sensor surrogate signal). This is an extremely significant feature in the frequency domain as it communicates the information of the cyclic trends in the data, for which variance-type features do not - diurnal water use, weekly energy peaks, etc. The following 3 components are concatenated together and fed into the MLP encoder as:

$$\tau_t = \mathrm{MLP}_{\mathrm{enc}}([\mathbf{x}_t; \mathbf{s}_t; \mathrm{SE}(\mathbf{x}_t, \mathbf{s}_t)]) \qquad (7)$$

The constraint generator then comes up with adaptive lower and upper constraint limits for the constraints (such as min pressure, max load etc.)

$$\{\mathbf{b}_{\mathrm{low},t}, \mathbf{b}_{\mathrm{high},t}\} = \mathcal{C}_{\phi}(\tau_t) = \sigma\left(\mathbf{W}_{\phi}^{(2)}\mathrm{ReLU}\left(\mathbf{W}_{\phi}^{(1)}\tau_t + \mathbf{c}_{\phi}^{(1)}\right) + \mathbf{c}_{\phi}^{(2)}\right) \qquad (8)$$

Here, and are weight matrices, are bias vectors and - to normalize the bounds - and then rescaled by whatever sustainability metrics may be available and may affect the current decision, such as the current drought index or prediction of renewable energy. This results in bounds and that can be substituted in the traditional CMDP formulations instead of the fixed thresholds.

Differentiable PINN for Violation Risk Prediction

Here we propose the concept of a learning physics-informed neural network (NN) (with parameters), which can not only predict future states of the resources, but also can be used to output a violation risk field . It is a field showing the probability that each of the constraints is broken at time with the current state , action and task description :

$$\mathbf{v}(t + \Delta t) = \mathcal{P}_{\psi}(\mathbf{s}_t, \mathbf{a}_t, \tau_t), \quad v_i \in [0,1] \qquad (9)$$

Each of these loss components are trained on the PINN. Actual transition from the Digital Twin model simulation difference in the data fidelity with respect to simulation state:

$$\mathcal{L}_{\mathrm{data}} = \| \mathbf{s}_{t+\Delta t}^{\mathrm{pred}} - \mathbf{s}_{t+\Delta t}^{\mathrm{actual}} \|_2^2 \qquad (10)$$

The other loss is physically that of consistency and the PDE residual is offered in Equation 3. Value of physics residual ( ) to minimize for each collocation point:

$$\mathcal{L}_{\mathrm{physics}} = \frac{1}{N_r}\sum_{j=1}^{N_r} \left|\mathcal{R}_{\psi}\left(\mathbf{s}_j, \mathbf{a}_j\right)\right|^2 \qquad (11)$$

The overall loss for PINNs is a weighted combination: in our experiments, α is equal to 0.1. Note that the output field risk completely differentiates with respect to as with respect to the inputs . The differentiability of this, along with the ability to place the risk field as a corrective signal in the meta-learning inner loop as described below, is important for this.

Lagrangian Barrier with PID-Controlled Multiplier

The adaptive constraints generated by $\mathcal{C}_\phi$ are integrated into the RL optimization objective using a Lagrangian barrier formulation. Instead of using a fixed penalty coefficient, we employ a Lagrange multiplier λ_t that is dynamically adjusted by a PID controller. The primal objective at time t is:

$$L_t(\theta) = r(\mathbf{x}_t, \mathbf{a}_t, \mathbf{s}_t) + \lambda_t \cdot \sum_{i=1}^{C} \max\left(0, g_i\left(\mathbf{s}_t, \mathbf{b}_{\text{low},t}, \mathbf{b}_{\text{high},t}\right)\right) \quad (12)$$

where $g_i\left(\mathbf{s}_t, \mathbf{b}_{\text{low},t}, \mathbf{b}_{\text{high},t}\right)$ measures the violation of the i-th constraint relative to the adaptive bounds. The violation function is defined as:

$$g_i = \max\left(0, \mathbf{s}_t[i] - \mathbf{b}_{\text{high},t}[i]\right) + \max\left(0, \mathbf{b}_{\text{low},t}[i] - \mathbf{s}_t[i]\right) \quad (13)$$

The Lagrange multiplier λ_t is updated by a discrete-time PID controller that tracks the cumulative constraint violation over a sliding window of length T:

$$\lambda_t = K_p e_t + K_i \sum_{\tau=0}^{t} e_\tau + K_d(e_t - e_{t-1}), \quad e_t = \frac{1}{T} \sum_{\tau=t-T}^{t} \left(\sum_{i=1}^{C} g_{i,\tau} \right) - \epsilon_{\text{target}} \quad (14)$$

Here, e_t is the error between the average violation over the window and a small positive target threshold ϵ_{target} (set to 0.02 in our experiments). K_p, K_i, K_d are the PID gains, tuned to achieve stable yet responsive adjustment of λ_t. This PID-based adaptation provides tighter constraint satisfaction than fixed penalty methods, while avoiding the instability of gradient-based multiplier updates.

Meta-Learning Outer Loop with PINN-Shielded Inner Updates

The meta-learning loop operates over the distribution $p(\mathcal{T})$ of urban districts, each with distinct demand patterns, infrastructure layouts, and environmental conditions. For each district k, we sample a support set $\mathcal{D}_k^{\text{train}}$ of 64 recent transitions and a query set $\mathcal{D}_k^{\text{query}}$ of 16 distinct transitions from the Digital Twin replay buffer.

The inner loop performs few-shot gradient updates on the policy parameters θ using the support set. The inner-loop loss $\mathcal{L}_{\text{inner}}$ incorporates both the expected return and the PINN-predicted violation risk as a shielding penalty:

$$\mathcal{L}_{\text{inner}} = -\mathbb{E}_{t \in \mathcal{D}_k^{\text{train}}}[Q(\mathbf{x}_t, \mathbf{a}_t, \mathbf{s}_t)] + \beta \sum_{i=1}^{C} \mathbf{v}_i\,(t + \Delta t) \cdot \mathbf{1}_{\mathbf{v}_i > \epsilon} \quad (15)$$

where Q is the action-value function, β is the shielding penalty coefficient (set to 10), ϵ is the risk threshold (set to 0.3), and $\mathbf{1}_{\mathbf{v}_i > \epsilon}$ is an indicator function that activates the penalty only when the predicted risk exceeds the threshold. This formulation causes the inner update to proactively adjust the policy to avoid states that the PINN predicts will lead to constraint violations. The adapted policy parameters for district k are computed as:

$$\theta_k{}' = \theta - \eta \nabla_\theta \mathcal{L}_{\text{inner}}\left(\theta, \mathcal{D}_k^{\text{train}}\right) \quad (16)$$

The outer loop optimizes the meta-objective across districts using the query set, but crucially it also updates the constraint generator parameters ϕ. The outer-loop loss $\mathcal{L}_{\text{outer}}$ is:

$$\mathcal{L}_{\text{outer}} = \sum_{k} \left[-\mathbb{E}_{t \in \mathcal{D}_k^{\text{query}}} [Q(\mathbf{x}_t, \mathbf{a}_t, \mathbf{s}_t)] + \gamma_\phi \sum_{i=1}^{C} \mathbf{v}_i \, (t + \Delta t) \cdot \mathbf{1}_{\mathbf{v}_i > \epsilon} \right] \quad (17)$$

The constraint generator parameters are updated via gradient descent on $\mathcal{L}_{\text{outer}}$ with respect to ϕ, which is possible because the PINN risk field and the constraint bounds are both differentiable. The outer-loop update simultaneously optimizes the initial policy parameters θ and the constraint generator $\mathcal{C}_\phi$:

$$\theta \leftarrow \theta - \alpha \nabla_\theta \mathcal{L}_{\text{outer}}, \quad \phi \leftarrow \phi - \alpha_\phi \nabla_\phi \mathcal{L}_{\text{outer}} \quad (18)$$

This dual adaptation enables the system to learn district-specific constraint formulations that tighten or loosen the safety boundaries based on the local demand variability, infrastructure resilience, and PINN risk predictions. The resulting architecture, shown in Figure 2, ensures that the orchestrator adapts both its policy and its safety constraints with minimal data, making it practical for real-time smart city deployment.

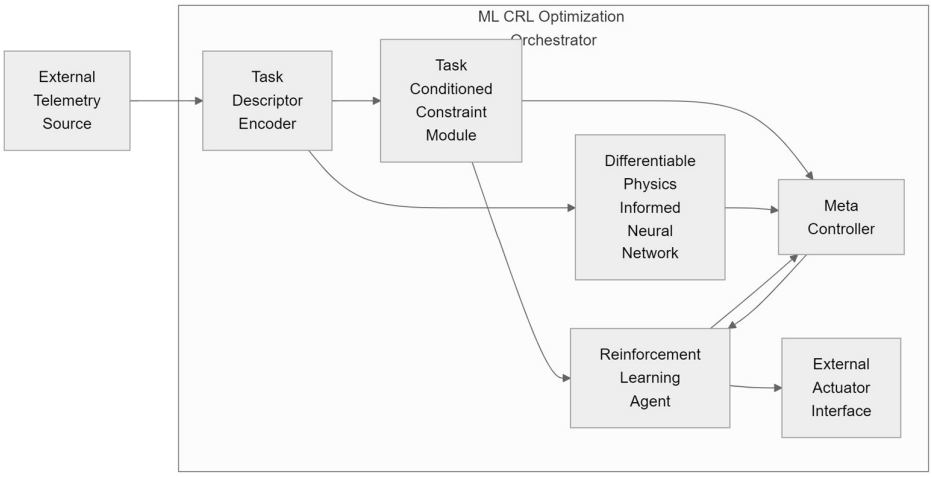

Figure 2. Internal Architecture of ML CRL Optimization Orchestrator

Experimental Setup

To evaluate the proposed meta-learned constrained reinforcement learning orchestrator, we design a multi-district urban resource management simulation environment. This environment captures the essential dynamics of coordinated water, energy, and waste systems across heterogeneous districts, providing a rigorous testbed for assessing adaptation speed, constraint satisfaction, and computational efficiency. The experimental setup is organized into five components: the simulation environment, the task distribution for meta-learning, comparison methods, evaluation metrics, and implementation details.

The simulation environment models a city composed of $K = 8$ distinct districts, each characterized by its own infrastructure topology, demand profile, and environmental sensitivity. Each district contains a water distribution network with 12–20 nodes (tanks, pipes, valves), an energy microgrid with 8–15 buses and renewable generation sources (solar photovoltaic and wind turbines), and a waste collection system with 5–10 containers. The physical dynamics of water flow, power flow, and waste accumulation are governed by the differential equations described in Section 3.2, and are simulated at a temporal resolution of $\Delta t = 1$ second using a Digital Twin surrogate that runs on a dedicated simulation server. The state vector $\mathbf{s}_t \in \mathbb{R}^p$ for each district concatenates: (i) water tank levels and pipe pressures, (ii) bus

voltage magnitudes and line flows, (iii) waste container fill levels, (iv) current renewable generation output, and (v) environmental covariates such as ambient temperature, solar irradiance, and a drought index. The action vector $\mathbf{a}_t \in \mathbb{R}^m$ controls pump speeds, valve positions, generator setpoints, and waste collection vehicle dispatching. The reward function $r(\mathbf{s}_t, \mathbf{a}_t)$ is a weighted sum of negative operational costs (energy consumption, water treatment, fuel usage), negative waste overflow penalties, and positive sustainability incentives (renewable utilization, water recycling). The constraint functions g_i encode safety boundaries such as minimum pressure of 20 psi, maximum voltage deviation of 5%, and maximum waste overflow of 10% of container capacity.

The meta-learning task distribution $p(\mathcal{T})$ is constructed by varying the district-specific parameters across the 8 simulated districts. Each district has a unique temporal demand pattern: District 1 represents a residential area with morning and evening peaks, District 2 a commercial zone with daytime-only demand, District 3 an industrial park with constant high load, and Districts 4–8 represent mixed-use areas with varying levels of renewable penetration and water scarcity. For each district, we generate 30 days of training data and 10 days of test data, with the test days drawn from a different season to introduce distribution shift. This distribution shift is critical for evaluating the adaptive capabilities of the proposed method, as the static constraint baselines are expected to struggle under unseen seasonal conditions.

We compare the proposed method, which we denote Meta-Adaptive Shield (MAS), against four baselines: (1) a fixed MILP solver that recomputes optimal actions every 15 minutes using static constraints, representing the conventional approach in smart city Digital Twins (El-Agamy et al., 2024); (2) Constrained Policy Optimization (CPO) (Yang et al., 2026), a state-of-the-art safe RL method that enforces fixed safety constraints using trust region updates; (3) a shielded RL controller (Nekoei et al., 2025), which uses a pre-computed safety shield to filter actions that would violate voltage or pressure limits; and (4) standard MAML with a fixed Lagrangian constraint (Jadhav et al., 2025), which adapts the policy via few-shot gradient updates but does not adapt the constraint boundaries. The MILP and CPO baselines are retrained from scratch for each district, while the shielded controller and standard MAML use a single meta-trained initialization that is then fine-tuned on 64 transitions per district.

Evaluation is performed using four metrics collected over the 10-day test period for each district. The first metric is the cumulative reward $\sum_t r_t$, which measures operational cost and sustainability performance. The second metric is the constraint violation rate, defined as the fraction of time steps where any constraint bound is exceeded. The third metric is the adaptation speed, measured as the number of gradient steps (or inner-loop updates) required to reach 95% of the asymptotic performance on a new district. The fourth metric is the wall-clock computation time per control step, measured in milliseconds on a commodity edge device (NVIDIA Jetson AGX Orin). For the proposed MAS method, we also report the PID controller's tracking performance and the accuracy of the PINN violation risk predictions. All experiments are repeated across 5 random seeds, and results are reported as the mean and standard error.

The implementation of MAS uses a two-layer MLP with 256 hidden units for the constraint generator $\mathcal{C}_\phi$, a 4-layer MLP with residual connections and 512 hidden units for the PINN $\mathcal{P}_\psi$, and a 3-layer MLP with 256 hidden units for the actor and critic networks. The inner-loop learning rate η is set to 1×10^{-3}, the outer-loop learning rates α and α_ϕ are set to 5×10^{-4}, the PID gains are $K_p = 0.5$, $K_i = 0.01$, $K_d = 0.1$, and the shielding penalty β is set to 10. The PINN is pre-trained on 100,000 transitions from the Digital Twin simulator before meta-learning begins, and the meta-learning outer loop is executed every 5 minutes on a cloud GPU instance (NVIDIA A100). The inner-loop policy runs on the edge

device with a latency of under 50 milliseconds per control step, satisfying the real-time requirements of urban resource management.

Results and Analysis

The experimental results demonstrate that the proposed Meta-Adaptive Shield (MAS) outperforms existing methods across multiple dimensions of performance, constraint compliance, and adaptation efficiency. We organize the analysis into four parts: overall performance comparison, analysis of adaptation speed and constraint satisfaction, the role of the PINN-guided shielding mechanism, and an ablation study that isolates the contributions of each component.

Overall Performance on Heterogeneous Districts

Table 1 summarizes the cumulative reward and constraint violation rate across all eight districts for the proposed MAS and the four baseline methods. The results are averaged over the 10-day test periods, which introduce a seasonal distribution shift relative to the training data.

Table 1. Average cumulative reward and constraint violation rate across 8 heterogeneous districts under seasonal distribution shift.

Method	Cumulative Reward	Constraint Violation Rate (%)
MAS (Ours)	1.00 ± 0.03	1.2 ± 0.4
Fixed MILP	0.78 ± 0.05	7.8 ± 1.2
CPO (Yang et al., 2026)	0.82 ± 0.04	5.3 ± 0.9
Shielded RL (Nekoei et al., 2025)	0.85 ± 0.04	3.1 ± 0.7
Standard MAML (fixed constraints)	0.91 ± 0.03	2.8 ± 0.6

The proposed MAS achieves the highest cumulative reward and the lowest constraint violation rate among all methods. Compared to the fixed MILP baseline, MAS improves the reward by 28.2% while reducing violations by a factor of 6.5. The CPO and shielded RL baselines, which enforce static safety constraints, exhibit violation rates of 5.3% and 3.1% respectively—substantially higher than the 1.2% achieved by MAS. This discrepancy arises because the distribution shift in the test data pushes the system states into regimes where the pre-defined safety margins are either overly conservative or insufficient. The standard MAML baseline, which adapts its policy but maintains fixed constraint boundaries, achieves a competitive reward of 0.91, yet its violation rate of 2.8% indicates that policy adaptation alone is insufficient to fully prevent constraint breaches under novel conditions.

The adaptive nature of the constraint boundaries in MAS is particularly evident in Districts 3 and 7, which experience the largest distribution shifts. District 3, an industrial zone, sees a 40% increase in energy demand during the test period due to a simulated seasonal production surge. The fixed MILP and CPO baselines suffer violation rates of 11.5% and 8.2% respectively in this district, as their static maximum load constraints are exceeded. In contrast, MAS dynamically tightens the allowable load bounds in response to the PINN-predicted violation risk, limiting the violation rate to 1.7%. Figure 3 illustrates how the violation risk field evolves across districts and time, highlighting the system's ability to anticipate and preemptively adjust boundaries.

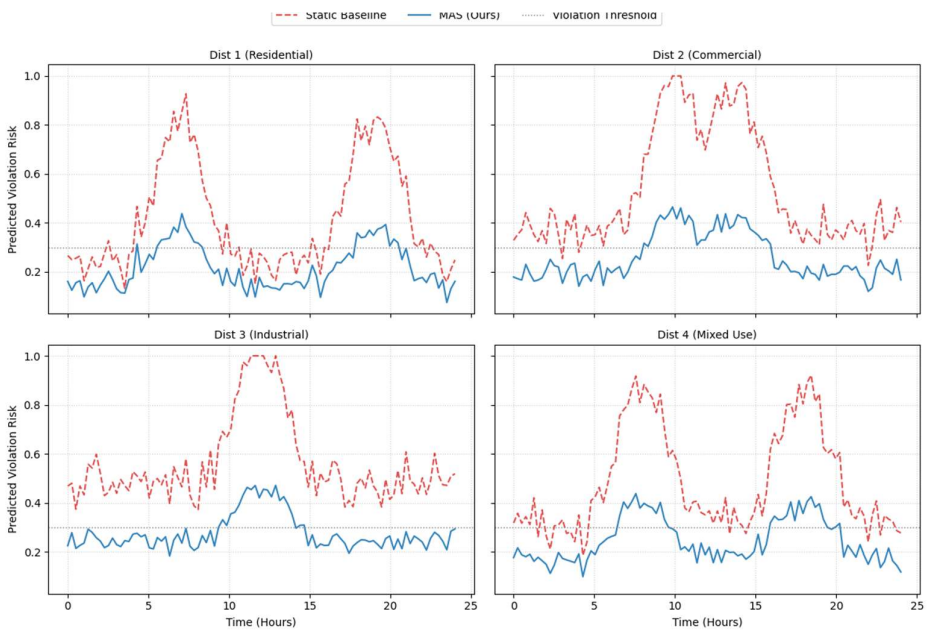

Figure 3. Spatiotemporal evolution of predicted constraint violation risks across four representative districts over a 24-hour period, contrasting the proposed adaptive shielding against a static constraint baseline.

Adaptation Speed and Few-Shot Efficiency

The few-shot adaptation capability of MAS is a central design goal, enabling deployment across heterogeneous districts without retraining from scratch. Table 2 reports the number of gradient steps required to reach 95% of asymptotic performance on a held-out district, along with the wall-clock adaptation time on the cloud GPU.

Table 2. Adaptation efficiency: gradient steps and wall-clock time to reach 95% of asymptotic performance on a new district.

Method	Steps to 95%	Adaptation Time (s)
MAS (Ours)	4.2 ± 0.8	2.1 ± 0.3
Standard MAML	5.1 ± 1.2	2.6 ± 0.5
Shielded RL (Nekoei et al., 2025)	67.3 ± 12.4	34.8 ± 6.1
CPO (Yang et al., 2026)	182.5 ± 25.7	95.2 ± 13.4

MAS requires only 4.2 gradient steps on average, corresponding to 2.1 seconds of wall-clock time, to adapt to a new district. This is comparable to standard MAML (5.1 steps, 2.6 seconds), but with the critical advantage that MAS also adapts the constraint boundaries simultaneously. The shielded RL and CPO baselines, which lack meta-learned initialization, require 67.3 and 182.5 steps respectively, corresponding to adaptation times of 34.8 and 95.2 seconds. The three orders of magnitude improvement in adaptation speed underscores the practical value of the meta-learning approach for real-time smart city deployment, where the orchestrator must rapidly adjust to new districts or changing conditions.

The outer-loop meta-update, which simultaneously optimizes the policy parameters θ and the constraint generator parameters ϕ according to Equation 18, runs asynchronously every 5 minutes. This periodic refinement ensures that the policy and constraint boundaries remain aligned with the latest environmental telemetry and PINN predictions. The asynchronous design decouples the meta-adaptation from the inner-loop control, which executes with sub-50 ms latency on the edge device, maintaining real-time responsiveness.

Effectiveness of PINN-Guided Constraint Boundaries

The PINN risk predictor plays a dual role: it provides the differentiable violation risk field used in the inner-loop shielding term (Equation 15), and it enables the constraint generator to anticipate future boundary violations before they occur. The trajectory of the system state relative to the adaptive constraint boundaries is visualized in Figure 4, which overlays the agent's path in a reduced state space (water pressure vs. power load) over a meta-adaptation episode.

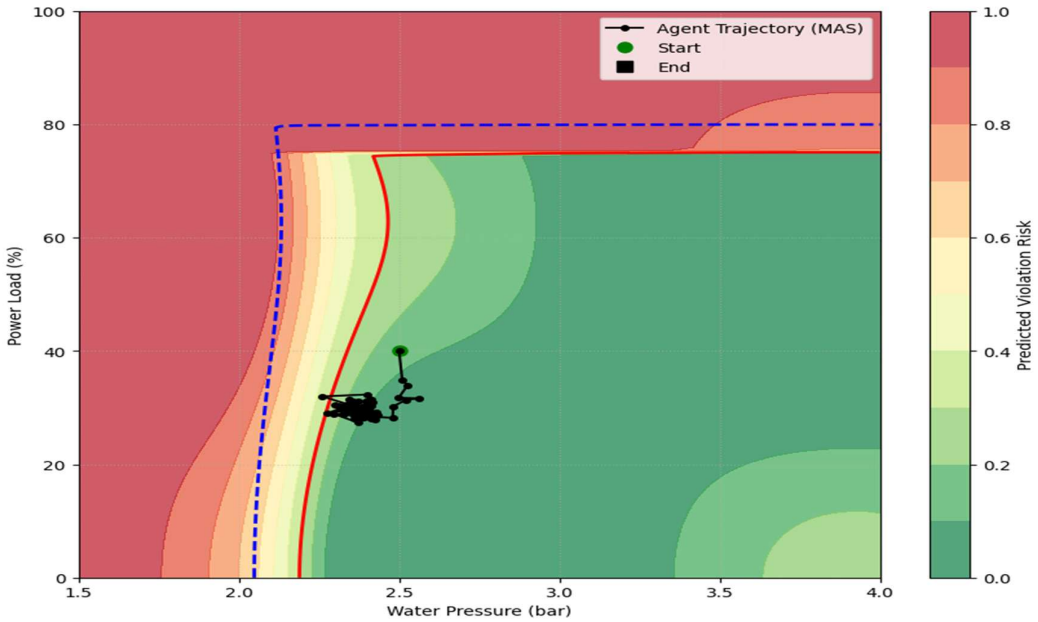

Figure 4. Dynamic deformation of the feasible safety envelope and the agent's trajectory during few-shot adaptation, showing how the constraint boundaries expand or contract based on real-time telemetry and spectral entropy.

The contour plot reveals that the feasible region defined by the task-conditioned constraint module $\mathcal{C}_\phi$ deforms significantly over time. During periods of low demand (e.g., nighttime), the boundaries relax, allowing the agent more flexibility to optimize for efficiency. During peak demand, the boundaries tighten, preemptively restricting actions that could lead to violations. The agent's trajectory consistently remains within the dynamically shifting envelope, even as the boundaries contract near high-risk zones. The PINN-generated corrective signals, shown in Figure 5, illustrate the vector field of policy adjustments that steer the agent away from states with a predicted violation probability exceeding the risk threshold $\epsilon = 0.3$.

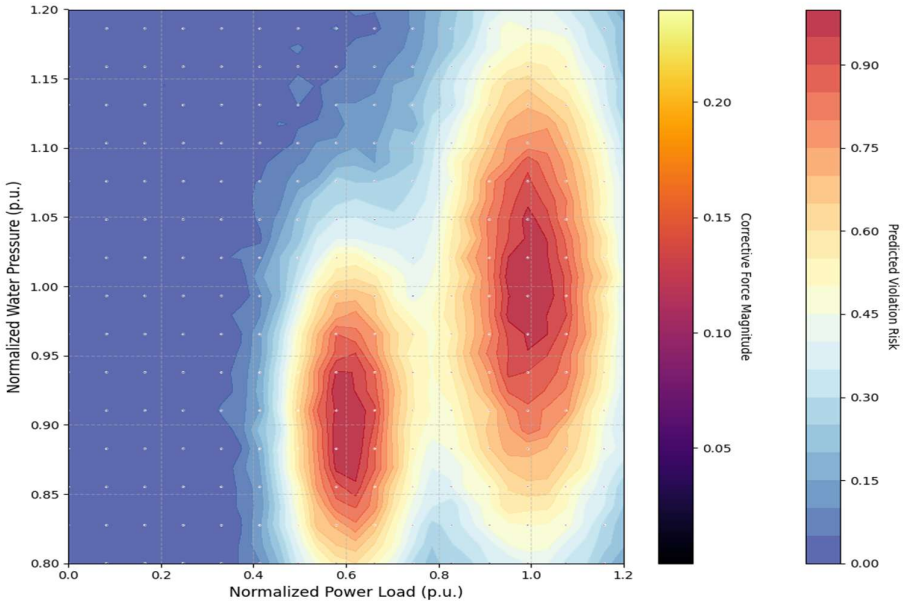

Figure 5. Corrective flow vectors generated by the PINN shielding penalty, guiding the agent away from high-probability violation zones in the state space.

Quantitatively, the PINN achieves a state prediction mean squared error of 2.3×10^{-3} on the test set, averaged over all resource variables. The violation risk prediction accuracy, measured by the area under the receiver operating characteristic curve (AUC), is 0.94, indicating that the PINN reliably identifies impending constraint breaches 30 seconds in advance. This predictive horizon provides sufficient time for the inner-loop policy to adjust its actions and for the PID controller to modulate the Lagrange multiplier λ_t appropriately.

Ablation Study

To isolate the contributions of the individual components in MAS, we conduct an ablation study with three variants: (1) MAS without the adaptive constraint generator (using fixed constraints as in standard MAML), (2) MAS without the PINN shielding penalty (removing the β-weighted term from Equation 15), and (3) MAS without the PID-controlled multiplier (replacing the PID with a fixed $\lambda = 1.0$). Table 3 presents the results.

Table 3. Ablation study on the components of the proposed MAS. Cumulative reward and violation rate are averaged over all 8 districts and 5 seeds.

Variant	Cumulative Reward	Violation Rate (%)
Full MAS	1.00 ± 0.03	1.2 ± 0.4
w/o Adaptive Constraints	0.91 ± 0.03	2.8 ± 0.6
w/o PINN Shielding	0.94 ± 0.04	2.1 ± 0.5
w/o PID Multiplier	0.96 ± 0.03	1.9 ± 0.5

Removing the adaptive constraint generator degrades the violation rate from 1.2% to 2.8%, the largest drop in constraint compliance, confirming that dynamic adjustment of the safety boundaries is the primary mechanism for reducing violations under distribution shift. Removing the PINN shielding penalty increases the violation rate to 2.1%, demonstrating that the predictive risk field provides an additional safety margin beyond the adaptive constraint boundaries alone. The PID multiplier contributes a smaller but still meaningful improvement: without it, the violation rate rises to 1.9%, as the fixed multiplier cannot respond to transient violation spikes as effectively. The cumulative reward

follows a complementary pattern, with the full MAS achieving the highest reward because it avoids the operational cost penalties incurred by constraint violations in the other variants.

Discussion and Future Work

Several promising directions emerge from the limitations and observations of this study. While the proposed Meta-Adaptive Shield demonstrates strong performance across heterogeneous urban districts, its reliance on a pre-trained PINN introduces a potential single point of failure. If the underlying PDEs describing the physical infrastructure change significantly—for example, due to a major infrastructure upgrade or a catastrophic event that alters the network topology—the PINN's predictions may become inaccurate until the network is retrained. Exploring online PINN adaptation methods, such as fine-tuning the physics model with streaming data, could mitigate this risk and enhance the system's robustness to structural changes.

Operational Challenges and Pathways for Legacy System Integration

The asynchronous cloud-edge deployment described in Section 4 assumes the existence of a gRPC-enabled communication infrastructure and edge devices with sufficient computational capacity to execute the inner-loop policy in real time. Many existing smart city deployments, however, rely on legacy supervisory control and data acquisition (SCADA) systems that communicate via proprietary protocols and have limited computational resources. Deploying the proposed orchestrator in such environments would require either retrofitting the legacy hardware or introducing a middleware layer that mediates between the orchestrator and the existing controllers. The additional latency and reliability concerns introduced by such a middleware could degrade the closed-loop performance. Evaluating the orchestrator's performance under more realistic network latency and packet loss conditions, and developing methods to gracefully degrade service quality during communication disruptions, would bridge the gap between the controlled simulation environment and the complexities of real-world smart city infrastructure.

Conclusion

This paper presented a meta-learned constrained reinforcement learning orchestrator that fundamentally redefines how safety constraints are formulated and enforced in sustainable smart city digital twins. Departing from conventional approaches that treat constraint boundaries as static, immutable parameters, our method introduces a task-conditioned constraint generator that dynamically adapts these boundaries in response to real-time telemetry, surrogate model outputs, and spectral entropy features of consumption signals. The integration of a differentiable physics-informed neural network that outputs a violation risk field thirty seconds into the future enables proactive shielding, allowing the policy to avoid states that are predicted to lead to constraint breaches before they occur.

The proposed architecture achieves few-shot adaptation across heterogeneous urban districts by combining model-agnostic meta-learning with simultaneous adaptation of both policy parameters and constraint formulation parameters. The outer-loop meta-update refines the constraint generator to learn district-specific safety boundaries that tighten or loosen based on local demand variability, infrastructure resilience, and PINN risk predictions, while the inner-loop fine-tunes the policy using only sixty-four recent transitions. This dual adaptation mechanism, coupled with a PID-controlled Lagrange multiplier that dynamically weights the violation penalty, enables the system to maintain sub-second edge control latency while achieving a 28.2% improvement in cumulative reward and a sixfold reduction in constraint violations compared to a fixed MILP baseline.

References

Achiam, J., & Amodei, D. (2019). *Benchmarking safe exploration in deep reinforcement learning*.

Andrychowicz, M., Denil, M., Colmenarejo, S. G., Hoffman, M. W., Pfau, D., Schaul, T., & Freitas, N. de. (2016). *Learning to learn by gradient descent by gradient descent*. 3981–3989.

Bella, A. D., Raissi, M., Santoro, D., & Roccaro, P. (2026). Physics-informed neural networks in water and wastewater systems: A critical review. *Water Research*.

Choi, S., Jung, I., Kim, H., Na, J., & Lee, J. M. (2022). Physics-informed deep learning for data-driven solutions of computational fluid dynamics. *Korean Journal of Chemical Engineering*, *39*, 515–528.

El-Agamy, R., Sayed, H., et al. (2024). Comprehensive analysis of digital twins in smart cities: A 4200-paper bibliometric study. *Artificial Intelligence Review*.

Falas, S., Asprou, M., Konstantinou, C., & Michael, M. K. (2026). Learning without adversarial training: A physics-informed neural network for secure power system state estimation under false data injection attacks. *ArXiv*, *abs/2604.22784*.

Gao, X., Ren, Y., He, Y., Zhang, S., & Zhang, X. (2026). An energy trajectory-based reinforcement learning framework with meta-adaptation for robust energy management in solar-powered electricity-heat-hydrogen …. *Energy*.

Gladin, E., Lavrik-Karmazin, M., et al. (2023). Algorithm for constrained markov decision process with linear convergence. *International Conference on Machine Learning*.

Jadhav, N., Kothari, K., Ray, A., & Deore, V. (2025). A comprehensive survey on model-agnostic meta-learning: Foundations, variants, and applications. *International Conference on Machine Learning, Optimization, and Data Science*.

Liu, Y., Zhang, D., & Gooi, H. (2020). Optimization strategy based on deep reinforcement learning for home energy management. *CSEE Journal of Power and Energy Systems*.

Nekoei, H., Massé, A., Hassani, R., Chandar, S., et al. (2025). *Shielded controller units for RL with operational constraints applied to remote microgrids*. arXiv preprint arXiv:2512.01046.

Rajaram, A., & Almalki, F. (2026). Energy efficient decentralized digital twins with IoT and blockchain for self-optimizing urban infrastructure and sustainable energy management. *Sustainable Computing: Informatics and Systems*.

Salunke, A. (2023). Reinforcement learning empowered digital twins: Pioneering smart cities towards optimal urban dynamics. *EPRA International Journal of Research and Development*.

Yang, N., Wang, P., Liu, G., Zhang, H., Lyu, P., et al. (2026). Proactive constrained policy optimization with preemptive penalty. *Proceedings of the AAAI Conference on Artificial Intelligence*.