Les Cahiers du GERAD ISSN: 0711–2440

Managing short-term hydropower production with a multi-agent reinforcement learning model

Y. Villeneuve, S. Séguin, M. F. Anjos, K. Demeester

G–2026–37

July 2026

La publication de ces rapports de recherche est rendue possible grâce au soutien de HEC Montréal, Polytechnique Montréal, Université McGill, Université du Québec à Montréal, ainsi que du Fonds de recherche du Québec – Nature et technologies.

Dépôt légal – Bibliothèque et Archives nationales du Québec, 2026
– Bibliothèque et Archives Canada, 2026

The publication of these research reports is made possible thanks to the support of HEC Montréal, Polytechnique Montréal, McGill University, Université du Québec à Montréal, as well as the Fonds de recherche du Québec – Nature et technologies.

Legal deposit – Bibliothèque et Archives nationales du Québec, 2026
– Library and Archives Canada, 2026

GERAD HEC Montréal
3000, chemin de la Côte-Sainte-Catherine
Montréal (Québec) Canada H3T 2A7

Tél. : 514 340-6053
Téléc. : 514 340-5665
info@gerad.ca
www.gerad.ca

Managing short-term hydropower production with a multi-agent reinforcement learning model

Yoan Villeneuve [a, b]

Sara Séguin [a, b]

Miguel F. Anjos [b, c]

Kenjy Demeester [d]

[a] *Université du Québec à Chicoutimi (Qc), Canada, G7H 2B1*

[b] *GERAD, Montréal (Qc), Canada, H3T 1J4*

[c] *School of Mathematics & Maxwell Institute for Mathematical Sciences, University of Edinburgh, Edinburgh, United Kingdom, EH9 3FD*

[d] *Rio Tinto Complexe Jonquière, Saguenay (Qc), Canada, G7S 4L2*

yoan.villeneuve1@uqac.ca
s1seguin@uqac.ca
miguel.f.anjos@ed.ac.uk
kenjy.demeester@riotinto.com

July 2026
Les Cahiers du GERAD
G–2026–37

Abstract : Hydropower generation fulfills 94% of Québec's electricity needs, making efficient Short-Term Hydropower Scheduling (STHS) critical for daily operations. This paper addresses the STHS problem using a multi-agent Reinforcement Learning (RL) model based on Proximal Policy Optimization (PPO). Multiple agents are trained to select hourly discharge actions defined by efficiency points, pairs of discharge and power output, while a reward function integrates reservoir management. A digital-twin of the hydropower system, based on twelve years of hourly operational data, serves as the environment. Model performance is evaluated over a 96-hour prediction horizon and validated out-of-sample against a Mixed-Integer Linear Programming (MILP) model. Results are analyzed through reservoir volume trajectories and energy production profiles, demonstrating the model's ability to generate operationally coherent schedules and highlighting the potential of advanced reinforcement learning approaches for improved hydropower decision-making.

Keywords: Reinforcement learning; short-term hydropower scheduling (STHS); hydropower optimization; proximal policy optimization (PPO); mixed-integer linear programming (MILP)

Résumé : L'hydroélectricité couvre 94% des besoins en électricité du Québec, ce qui rend la planification à court terme de la production hydroélectrique (PHCT) essentielle à son fonctionnement quotidien. Cet article aborde le problème de la PHCT à l'aide d'un modèle d'apprentissage par renforcement multi-agents basé sur l'optimisation de politique proximale (OPP). Plusieurs agents sont entraînés à sélectionner des actions de débit horaires définies par des points d'efficacité, des paires débit-puissance, tandis qu'une fonction de récompense intègre la gestion du réservoir. Un jumeau numérique du système hydroélectrique, basé sur douze années de données d'exploitation horaires, sert d'environnement. La performance du modèle est évaluée sur un horizon de prédiction de 96 heures et validée hors échantillon par rapport à un modèle de programmation linéaire en nombres entiers mixtes (PLNE). Les résultats sont analysés à travers les trajectoires de volume du réservoir et les profils de production d'énergie, démontrant la capacité du modèle à générer des calendriers opérationnels cohérents et soulignant le potentiel des approches d'apprentissage par renforcement avancées pour une meilleure prise de décision en matière d'hydroélectricité.

Mots clés : Apprentissage par renforcement; planification de la production hydroélectrique à court terme (PHCT); optimisation de l'énergie hydroélectrique; optimisation de politique proximale (OPP); programmation linéaire en nombres entiers mixtes (PLNE)

1 Introduction

Hydropower management is a complex decision-making process, as the environment that surrounds a hydropower system is filled with uncertainty and production constraints [20]. In a regulated market with fixed price, efficiency of energy production and resource management is most often prioritized. The STHS problem focuses on finding the optimal quantity of water to be used at hourly or daily intervals, over a specific length of time and with consideration to the unit assignment in the powerhouse [16]. The literature contains a wide range of algorithms to solve the STHS problem, like dynamic programming optimization [13], mix integer programming model [9, 12, 21] and machine learning [22, 25]. With the rapid advancement seen in the development of machine learning algorithms [5], previous works [24] explored the use of a Recurrent Neural Network (RNN) with a Long Short-Term Memory (LSTM) layer to determine if it could be trained with a historical dataset to make hourly predictions over a sequence of 96 hours. It was concluded that it is possible, as results show potential when compared to a MILP model. This paper addresses the impact of the predictive model on the hydropower system and to introduce a new model based on RL. This type of model allows for a better representation of the hydropower plants, as RL enables a more flexible approach than the purely data driven training of conventional RNN by integration of the reward function as a means to guide the prediction behaviour and measure the performance. There exist many types of RL methods that treat different problem application. The choice depends mainly on the type of problem to solve, the specificity of the environment and how the model is supposed to learn [1]. This study settled on PPO, a model-free and on-policy algorithms for RL [19]. This method is characterized by updates during training that directly affect the policy used to make decisions. The distinctive feature of PPO is that the magnitude of each policy update is constrained, which helps prevent performance collapse, but reduces training efficiency.

The motivation to work on this paper is inspired by [15], a paper that explores the medium term problem of the water resource management for multiple reservoirs. A policy gradient method called *REINFORCE* [27] is used with chance constraints to handle the water storage. This paper also introduces a novel approach called *backoffs*, which helps with the implementation of the chance constraints. Results show superior efficiency of resource management on multiple reservoirs and good scaleability. In [18], a RL model is used to demonstrate its potential for long-term hydropower scheduling to maximize annual profit under weekly inflows and electricity prices, though validated only on a simplified case study.

To our knowledge, few papers use RL to solve the STHS problem, and none has attempted a multi-agent approach for the operation of a hydropower system. In the short-term, [7] used a soft actor-critic algorithm for continuous hourly operation of an interconnected hydropower system in Ecuador, with promising results compared to historical equivalent. In [8], three RL methods, Advantage Actor-Critic, PPO and soft actor-critic were tested on a two reservoir system and a simulated six-reservoir system for intraday operations. Benchmark was done with a MILP model. Finding show that MILP outperformed the RL algorithm on the two reservoir system, but PPO scaled better over the six reservoir system. In [26], a Multi-stage RL algorithm with a random environment simulation, combine with an optimization model, achieve higher efficiency in a cascade hydropower environment using water levels as the predicted action of the model.

This paper aims to develop a RL model to address the STHS problem of hourly operation in a system of interconnected hydropower plant. The model can predict a series of 96 hourly decisions base on previous environmental states. The model is based on the Independent PPO (IPPO) algorithm to build independent predictors for each hydropower plant, while the impact on the system is shared across the model. Further consideration is implied through the environment, like the efficiency of the production through efficiency point (EP) [9] and unit management. The hyperparameters of the model have been selected through a black box optimization solver called Nomad [2] and a MILP model is used as a benchmark for the results.

2 Case study

The targeted environment plays an important role and the reproducibility of a RL model. In hydropower, there are a number of factors that can influence the daily operational decisions of a plant. The location, climate and wildlife all carry sets of constraints that must be considered. Human activities can also be of great importance, like residence, leisure or other infrastructure that may rely on the reservoir level for socio-economic activities. Furthermore, the electricity market and demands often differ from one country to another. The hydropower system environment in this paper is based in the region of Saguenay Lac-Saint-Jean in the province of Québec, Canada. The system is prone to seasonal changes with low inflow period in the winter and high natural inflow in spring. It is made up of two interconnected hydropower plants named Chute-du-Diable (CD) and Chute-à-la-Savane (CS) on the Peribonka River. Both plants are separated by about $20km$ of river length, with CD located upstream. Given the proximity between plants, operational decisions are closely related and the travel time of water is not considered.

These hydropower plants are part of a larger system operated by Rio Tinto, an aluminum producer that operates many factories in this region, which covers about 90% of their energy needs. Therefore, the electricity market is not considered, as production is meant to be fully exploited for aluminum production and the remainder is bought from Hydro-Québec through fixed-price contracts. A dataset of hourly production data from December 2010 to November 2022 has been made available to develop the model in this paper. Eight out-of-sample validation instances from years 2020 and 2021 are also available to test the model. Both plants share similar configurations, each consisting of 5 generating units. CD has a dam height of $37.8m$ and a gross water head of $33.1m$, while CS has a dam height of $39.62m$ and a gross water head of $34.5m$. A key difference between these plants is the storage size, where CD has a reservoir capacity of $1200hm^3$ compared to $625hm^3$ for CS. The installed capacity of CD is $224MW$ with a maximum discharge of $850m^3/s$, whereas CS has a slightly higher installed capacity of $245MW$ and a maximum discharge of $810m^3/s$ [14, 23].

3 Methodology

A RL algorithm is implemented based on the interactions between the agent and the mutable environment through each step. Therefore, a digital-twin of the hydropower system is developed based on true historical data. A multi-agent PPO model is introduced in the environment to manage the hourly discharge value for each hydropower plant. Most RL algorithms operate within the Markov Decision Process (MDP) framework, which relies on the tuple SARS (State, Action, Reward, next State) to describe the agent and environment interaction [11]. Therefore, the model description is split into four sub-sections, the environment, the action, the agent and the transition step. An overview of the IPPO model introduced in this section is presented in Figure 1 based on the MDP framework, which showcases the SARS interaction between the agents and the environment.

Each agent receives a state S_t that represents the observable environment for a given period t and takes an action A_t based on the given state S_t using its own decision policy π. This action affects the current state of the environment, which is represented by the new state S_{t+1}. The transition to the next state generates a reward R_t that computes the positive (or negative) impact of the transition $S_t \rightarrow S_{t+1}$. The cycle continues as the new state is fed to the agent until an end condition is met. During the learning process, the agent periodically receives pairs of next states and rewards to update its own decision policy. This ensures that the policy always makes decisions that return the largest cumulative reward. The sub-sections that follow contextualize this framework to the current problem.

3.1 Environment

Based on the available data, the environment is constructed so that each state S_t for both plants represent the production at a given hour of a given date. Because of the winter, the months of March

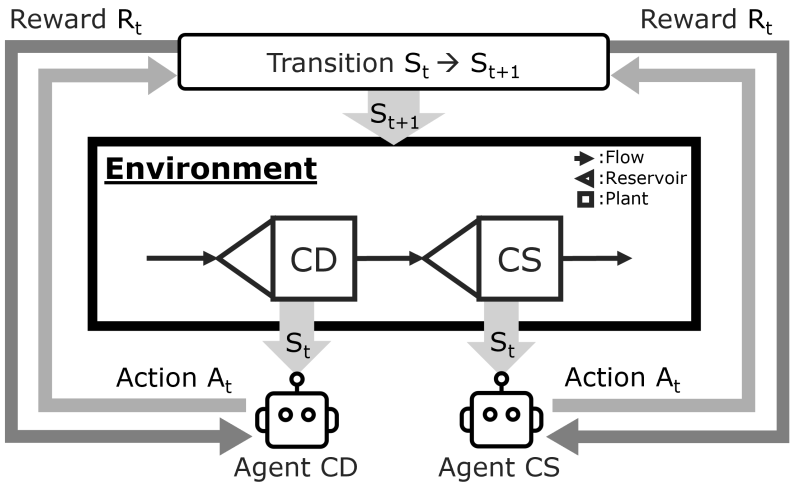

Figure 1: Representation of the IPPO model interaction between the agents and the hydropower environment.

until June were removed for the dataset, as these periods are characterized by high amounts of natural inflow cause by the snowmelt. The context is so different from the rest of the year that it creates bias when the model is trained. The dataset gives insight on the production value and hourly states of the powerhouses and reservoirs. These features are used to reconstruct the environment such that each timestep can transition with the historical sequence. Table 1 presents the features that defines each state S_t of the hydropower system.

Table 1: Environment features available and observable by agents for both powerplants, except for *, available only for CS. Initial and final volumes are imposed.

Feature	Unit	Feature	Unit
Timestep (T)	[0, 95]	Prev. Discharge	m^3/s
Week	[1, 52]	Natural Inflow	m^3/s
Reservoir Volume (v_t)	hm^3	Active units	[3, 5]
Initial Volume (V_i)	hm^3	Upstream discharge*	m^3/s
Final Volume (V_f)	hm^3		

Each feature represents a state value for a given hydropower plant, meaning that the agents only see their assigned powerplant features. Furthermore, the upstream discharge and spill are features only present for the CS agent. These represent the discharge decision and planned spill for upstream plant CD. The initial and final volume are the only values that stays fix for the duration of the prediction and are also present in the benchmark MILP model [9]. More context for these features is presented in Section 3.4.

3.2 Action

The set of actions A that an agent can take is based on EPs, introduced in [9]. These are pairs of fix water discharge and energy that are already known to be the best value to operate a given hydropower plant for a given combination of active turbines. Therefore, the set of actions A represents every discharge values present in the set of EPs. For better control of the reservoir, more EPs have been added in the set base on the adjacent discharge values of EP already present. Using predetermined discharge value to discretize the action set A greatly reduces the amount of possible outcome. Base on the current state of the hydropower plant, the environment assign each plant to a given EP in the transition step.

3.3 Agent

With an environment composed of two hydropower plant, the model requires more than a single decision maker to operate each hydropower plant, because each plant has their own set of actions A and has their own impact on the environment. This is referred to as multi-agent RL. There are two general ways to implement multiple agents in PPO, either with IPPO, where each agent is trained independently, or Multi-Agent Deep Reinforcement Learning (MAPPO), which relies on centralized training with decentralized execution. In an IPPO model, the agents must learn to cooperate together in order to maximize the cumulative reward, which adds complexity to the learning process. In MAPPO, each agent has an actor network, but all share a single critic. This help to take into consideration the whole environment and updates the model accordingly. In this paper, both methods have been implemented and tested, but IPPO was chosen because of its superior performance, which seems to align with [10].

3.4 Transition step

The environment is a dynamic model designed to simulate the evolution of the hydropower system through each timestep t until the end condition $t = T$ is reached, where $T = 96$. When an agent makes a decision, the impact is applied to the environment with the other external factors of the system to advance to the next state S_{t+1}. First, the impact on the reservoir volume of the hydropower plant is computed with the discharge action A_t, the natural inflow. Second, the action is associated with an EP to update the active units of the powerhouse and the energy produced accordingly. Because two EPs can share the same discharge value, the environment determines, based on the available turbine units, potential energy production and feasibility, the current EP and if a turbine unit should be activated or deactivated. The energy is also adjusted to the current reservoir volume. When the new state is available, the reward Eq.(3) is computed using Eq.(1) and Eq.(2).

$$r_i = \max\left(0,\ 1 - \frac{|V_i - v_t|}{\Delta v_{\max}}\right) \frac{T-t}{T}, \tag{1}$$

$$r_f = \max\left(0,\ 1 - \frac{|V_f - v_t|}{\Delta v_{\max}}\right) \frac{t+1}{T}, \tag{2}$$

$$R_t = \begin{cases} -1 & , v_t < v_{\min} \text{ or } v_t > v_{\max}, \\ 1 + \mathrm{r}_i + \mathrm{r}_f & , q_t \neq q_{t-1}, \\ 2 + \mathrm{r}_i + \mathrm{r}_f & , q_t = q_{t-1}. \end{cases} \tag{3}$$

The reward R_t focuses on the reservoir volume of a single plant. At first, a fixed penalty (negative reward) of -1 is returned if the volume does not respect the minimum or maximum bounds of the hydropower plant reservoir. If theses conditions are met, a base reward of one is given if the discharge q_t is different from the last discharge q_{t-1}. Ideally, the operator of a plant prefers to keep a stable discharge value for extended amounts of time to prevent wear and limit the activation of turbine units. If the discharge q_t stay the same, a base reward of two is given. Then, an additional reward from r_i and r_f, ranging from zero to 1 is added. These consider the closeness of the reservoir volume to either the initial volume V_i and the final volume V_f base on the current period t. When t is low, the gain from r_i is better. As t gets closer to the end condition $t = T$, the gain from r_f gets better. The range of proximity to the targeted volume is given by $\Delta v_{\max}$, which is a fixed parameter for each plant based on the capacity of the reservoir. The reward that is recorded during the training and testing is the sum of the cumulative reward from each hydropower plant.

4 Results and discussion

This section discusses the results for the IPPO model presented in the previous section. Hyperparameters of the model are also presented have they have been through and optimization step. A validation set of eight test instances is used to compare the IPPO model to a MILP model and historical decisions. The MILP model has been developed prior to this paper on the same hydropower system [9]. The problem solved in this paper is deterministic. Therefore, the same final volume as in the MILP is used as a final volume value for the IPPO.

The dataset was divided into train and test subsets, ranging from 2010 to 2020 for training and 2020 to 2022 for testing. A set of eight validation instances was also provided to benchmark the model with a MILP model. A *Nvidia H100 80GB HBM3* graphic card on Rorqual, a supercomputer operated by Calcul Québec [6], is used to train the model. The model was developed using PyTorch [17], a well-known Python library for machine learning. A hyperparameter search has been done in order to maximize training efficiency and performance of the model with Nomad [2], a black box optimization solver that uses the Mesh Adaptive Direct Search (MADS) algorithm [3] to optimize difficult problems. RL algorithms are very sensitive to changes in the algorithm parameters, which is why MADS has been applied to find the best configuration possible. Table 2 shows the values chosen for the IPPO algorithm.

Table 2: Hyperparameter configuration of the IPPO training procedure.

Hyperparameter	Value	Hyperparameter	Value
Learning rate (α)	5e-4	Entropy coefficient	0.10
Discount factor (γ)	0.77	Value loss coefficient	0.5
GAE parameter (λ)	0.28	Decay rate	0.94
Clipping parameter (ϵ)	0.30	K epochs	25
Mini-batch size	32	Rollout size	$20 \times T$
Input sequence length (T)	4		

The architecture of the actor and critic neural networks have also been optimized trough MADS. Given the high amount of customization available for neural networks, layer types, activation function where decided prior to the search. Base on previous findings [24], it has been found that RNNs are well suited to detect patterns that leads to improved predictions. The use of an LSTM layer after the input allows for multiple states instances, which was optimized to an input sequence length of four timesteps. Table 3 summarizes the architectures after the MADS on the number of dense layers, the number of neurons per layer and the amount of regularization with a dropout algorithm [4].

Table 3: Actor-Critic RNN architectures.

(a) Actor network

Layer #	Layer Type	# neurons	Activation	Dropout (%)
1	Input	d_{in}	–	–
2	LSTM	192	tanh	20
3,4	Dense	64	ReLU	–
5	Output	d_{out}	Softmax	–

(b) Critic network

Layer #	Layer Type	# neurons	Activation	Dropout (%)
1	Input	d_{in}	–	–
2	LSTM	192	tanh	20
3,4,5,6	Dense	192	ReLU	–
7	Output	1	Linear	–

The number of d_{in} input units depends on the size of the hydropower plant state, available in Table 1, multiplied by the input sequence length. The critic network of Table 3b is much larger and wider than expected compared to the actor network of Table 3a. The output d_{out} for the actor network is set to the number of actions A and a single linear value for the critic. A regularization is applied in the form of a dropout before the output layer, which randomly deactivates a subset of neurons during training [4]. After three hours of training, the model converges to an optimum, where further updates do not yield higher sums of rewards. The results of the model on the test set are presented in Figure 2.

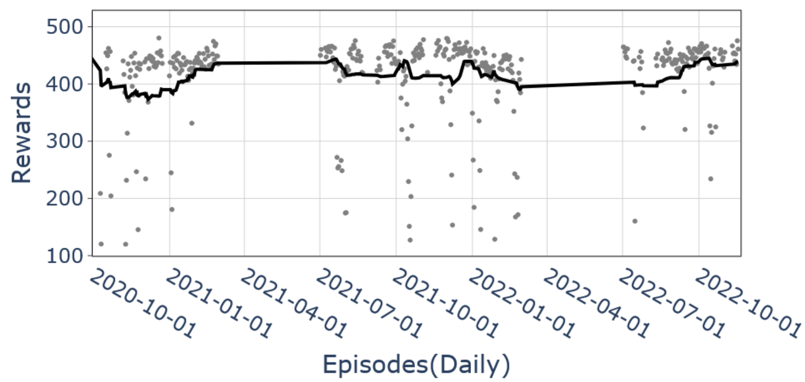

Figure 2: Reward obtained by testing each day available in the test set, with a moving average of the last 50 days.

Results from the training and testing step show similar performance in reward gain per instance, which imply it did not overfitted. There are stretches without points in the figure is caused by the removed months of snowmelt periods in the dataset. Some episodes have noticeable bad reward values because some instances cannot provide states that can satisfy operational constraints, such as the initial volume in the reservoir bounds.

The data provider does not allow the true values of the dataset to be disclosed in this manuscript, but can be converted into percentages with the Eq.(4). This equation is only used to display data in this paper, but never applied in the main project.

$$\% = \frac{current\ value - \min(histo.\ \ value)}{\max(histo.\ \ value) - \min(histo.\ \ value)} \times 100. \tag{4}$$

The runtime of the model on each validation instance takes about one second for a complete prediction. Figure 3 presents the percentage of energy produced in the validation instances of the IPPO model compared to historical production and the results of the MILP model, with the percentage of improvement of the IPPO model compared to the historical production.

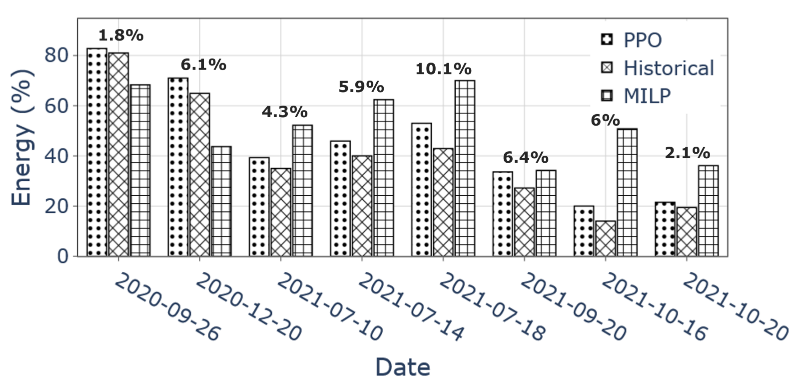

Figure 3: Results comparison between the IPPO prediction, historical production and MILP model. Improvement of the IPPO model over the historical decision is shown in percentages above each instance.

The IPPO model presents an improvement in every validation instances tested over the historical decision, with an average of 3.34% over the set. In all instances, operational constraints are respected. The MILP performed better than the IPPO, but from September 2020, the IPPO model did perform better. The year 2020 is marked by high volumes of natural inflows, which could indicate that the IPPO model performs better with high uncertainty. Activation of turbine units are kept to a minimum by the IPPO model, but some instances do reach above the MILP model. As noted in Section 2, CD and CS differ significantly in reservoir capacity. This disparity makes the operation of CD more complex than CS, as its larger reservoir leaves a lot more freedom to fluctuate. Consequently, the agent assigned to CD exhibited an interesting behaviour that can be observed in Figure 4, which presents the resulting volume of the IPPO model decisions for the instance of June 10, 2021.

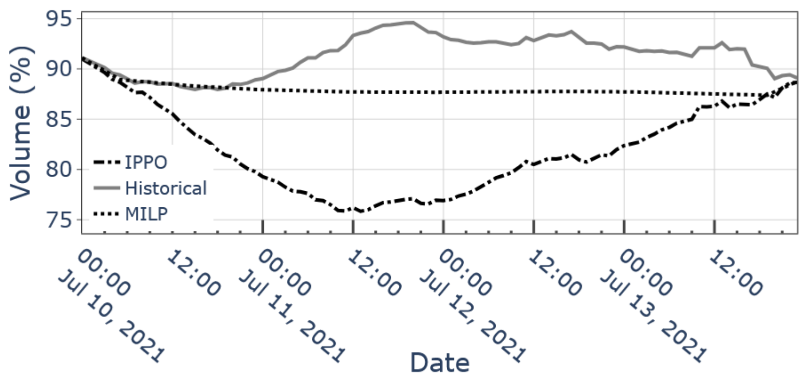

Figure 4: Comparison of the reservoir volume after the discharge is applied to historical variation of the CD hydropower plant.

For half of the validation instances, the reservoir volume of CD for the IPPO model results in a sequence where the mid-points reach a lower than expected volume, while still ending up close to the final volume (V_f). This behaviour is probably a method of risk aversion to prevent leaving the upper bound of the reservoir. The IPPO model seems to have a low confident in its ability to keep the reservoir at high levels, which would increase its cumulative rewards, but also risk gaining negative rewards based on the Eq. (3). Given that the input contains only data from previous states, it seems enough to reach a decision policy that is able to respect the operational constraints, but not to a degree where the model is confident enough to stay at the limits of the high volume bounds.

In light of these conclusions, future work should focus on implementing future natural inflow predictions data in the hydropower plant environment, better turbine unit management and energy management as a reward factor.

5 Conclusion

In this paper, an IPPO model is developed to address the STHS problem. An environment based on historical data of a system of two interconnected hydropower plants is created to allow multiple agents, one per hydropower plant, to predict a discharge sequence over a horizon of 96 hours.

Previous RNN models developed on the same dataset could not capture the operational dependency between predicted discharge and the reservoir volume, but the implementation of a RL algorithm with a reservoir based reward addressed this dependency. The hyperparameters are optimized with MADS and the model is trained on 12 years of historical data from 2010 to 2022. Although the IPPO model currently uses the last four production states to make predictions, an improvement of around 3.34% over the historical decision is observed, which is very promising. The model did not perform as well as the MILP model in the benchmark, but prediction times are near instantaneous and the model could scale to handle larger systems. Turbine unit management was implemented in relation with EPs in the environment, but the hydropower unit commitment problem could be further defined in future work. The reward function could also be improved with regards to energy production and natural inflows.

References

[1] Joshua Achiam. Spinning Up in Deep Reinforcement Learning. 2018.

[2] C. Audet, S. Le Digabel, V. Rochon Montplaisir, and C. Tribes. Algorithm 1027: NOMAD version 4: Nonlinear optimization with the MADS algorithm. ACM Transactions on Mathematical Software, 48(3):35:1–35:22, 2022.

[3] Charles Audet and J. E. Dennis. Mesh adaptive direct search algorithms for constrained optimization. SIAM Journal on Optimization, 17(1):188–217, 2006.

[4] Pierre Baldi and Peter J Sadowski. Understanding dropout. Advances in neural information processing systems, 26, 2013.

[5] Chiara Bordin, Hans Ivar Skjelbred, Jiehong Kong, and Zhirong Yang. Machine learning for hydropower scheduling: State of the art and future research directions. Procedia Computer Science, 176:1659–1668, 2020. Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 24th International Conference KES2020.

[6] Calcul Québec. calculquebec.ca. https://www.calculquebec.ca/, 2024. [Accessed 03-03-2026].

[7] José María Campo Carrera and Angel Udias. Deep reinforcement learning for complex hydropower management: evaluating soft actor-critic with a learned system dynamics model. Frontiers in Water, 7:1649284, 2025.

[8] Rodrigo Castro-Freibott, Álvaro García-Sánchez, Francisco Espiga-Fernández, and Guillermo González-Santander de la Cruz. Deep reinforcement learning for intraday multireservoir hydropower management. Mathematics, 13(1), 2025.

[9] Maissa Daadaa, Sara Séguin, Kenjy Demeester, and Miguel F Anjos. An optimization model to maximize energy generation in short-term hydropower unit commitment using efficiency points. International Journal of Electrical Power & Energy Systems, 125:106419, 2021.

[10] Christian Schröder de Witt, Tarun Gupta, Denys Makoviichuk, Viktor Makoviychuk, Philip H. S. Torr, Mingfei Sun, and Shimon Whiteson. Is independent learning all you need in the starcraft multi-agent challenge? CoRR, abs/2011.09533, 2020.

[11] Frédérick Garcia and Emmanuel Rachelson. Markov Decision Processes, chapter 1, pages 1–38. John Wiley & Sons, Ltd, 2013.

[12] Francesco Gerini, Elena Vagnoni, Rachid Cherkaoui, and Mario Paolone. Optimal short-term dispatch of pumped-storage hydropower plants including hydraulic short circuit. IEEE Transactions on Power Systems, 40(3):2050–2060, 2025.

[13] Yueyang Ji and Hua Wei. An approximate dynamic programming method for unit-based small hydropower scheduling. Frontiers in Energy Research, Volume 10 – 2022, 2022.

[14] Ministère de l'Environnement, de la Lutte contre les changements climatiques, de la Faune et des Parcs. Répertoire des barrages. https://www.cehq.gouv.qc.ca/barrages/listebarrages.asp, May 2018.

[15] Florian Mitjana, Michel Denault, and Kenjy Demeester. Managing chance-constrained hydropower with reinforcement learning and backoffs. Advances in Water Resources, 169:104308, 2022.

[16] Prahlad Mundotiya, Parul Mathuria, and Harpal Tiwari. Mathematical approach-based power system analysis: A review of short-term hydro scheduling. Intelligent Data Analytics for Power and Energy Systems, pages 85–98, 2022.

[17] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library, 2019.

[18] Signe Riemer-Sørensen and Gjert H. Rosenlund. Deep reinforcement learning for long term hydropower production scheduling. In 2020 International Conference on Smart Energy Systems and Technologies (SEST), pages 1–6, 2020.

[19] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. CoRR, abs/1707.06347, 2017.

[20] Sara Séguin. Hydropower optimization. Les Cahiers du GERAD ISSN, 711:2440, 2020.

[21] Hans Ivar Skjelbred, Jiehong Kong, and Olav Bjarte Fosso. Dynamic incorporation of nonlinearity into milp formulation for short-term hydro scheduling. International Journal of Electrical Power & Energy Systems, 116:105530, 2020.

[22] M. Thirunavukkarasu, Yashwant Sawle, and Himadri Lala. A comprehensive review on optimization of hybrid renewable energy systems using various optimization techniques. Renewable and Sustainable Energy Reviews, 176:113192, 2023.

[23] Rio Tinto. Installations. https://energie.riotinto.com/energie-electrique/installations/, Mar 2021.

[24] Y Villeneuve, S Séguin, A Chehri, and K Demeester. Short-term hourly hydropower prediction: Evaluating lstm and milp-based methods. Les Cahiers du GERAD ISSN, 711:2440, 2025.

[25] Yoan Villeneuve, Sara Séguin, and Abdellah Chehri. Ai-based scheduling models, optimization, and prediction for hydropower generation: Opportunities, issues, and future directions. Energies, 16(8), 2023.

[26] Xin Wan, Xiaohui Yuan, and Zhiqiang Jiang. Random environment simulation-based multi-stage reinforcement learning for short-term scheduling of cascade hydropower stations. Energy, 345:140208, 2026.

[27] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. Machine learning, 8(3):229–256, 1992.