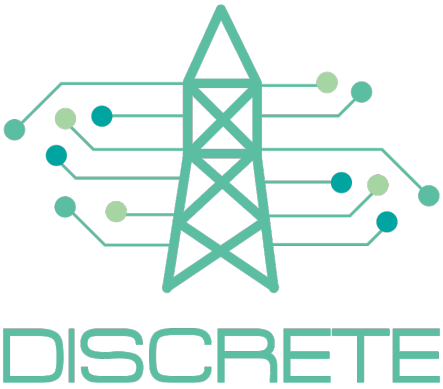

D2.4. Combining supervised learners (in D2.2) and the reinforcement learning model (in D2.3) to generate an improved model for the real-time operation of modern systems

Author: Jean-François Toubeau

This deliverable is based on the publication:

C. Rasic, P. Favaro, Y. Wang and J. -F. Toubeau, (2026) "Safe Reinforcement Learning for Battery Energy Storage Participation in the Imbalance Settlement," in IEEE Transactions on Energy Markets, Policy and Regulation, ***doi: 10.1109/TEMPR.2025.3639758****.*

Table of Contents

1. Purpose within the project

This work was carried out to investigate how supervised learning and reinforcement learning can be integrated within a unified decision-support framework for modern power-system operation. Earlier work in DISCRETE identified the most suitable supervised-learning models for optimisation problems and demonstrated how machine learning can accelerate optimisation procedures. D2.3 subsequently established the feasibility of autonomous decision-making through reinforcement learning. The objective of D2.4 is therefore to combine the complementary strengths of both paradigms within a single architecture capable of supporting real-time operational decision making.

The motivation is that supervised learning and reinforcement learning address different aspects of the decision-making problem. Supervised learning excels at extracting information from historical datasets, generating forecasts, and learning structured representations of system behaviour. Reinforcement learning, in contrast, is specifically designed to learn sequential decision-making strategies under uncertainty. Combining the two approaches therefore offers the possibility of creating more informed, robust, and adaptive operational policies.

Aspect	Summary
Objective	Combine supervised learning and reinforcement learning
Supervised component	Forecasting and expert-guided learning
RL component	Safe autonomous decision making
Main output	Real-time decision-support tool for operational planning and control
Expected benefits	Improved adaptability, computational efficiency, and operational performance

2. Methodological contribution

The proposed framework combines three complementary learning layers.

The first layer consists of supervised-learning models used to construct informative state representations. Forecasting models generate probabilistic predictions of future imbalance conditions and price distributions. These forecasts provide the RL agent with information regarding expected future system evolution and therefore improve its ability to anticipate future operating opportunities. Rather than reacting only to current conditions, the agent can incorporate forecast information directly into its decision-making process.

The second layer consists of supervised expert-guided learning. Prior to reinforcement-learning training, the framework generates offline datasets containing expert trajectories derived from heuristic strategies and simplified optimisation procedures. These trajectories are used to initialise the replay buffer and guide early exploration. This supervised pre-training phase prevents the RL agent from becoming trapped in poor local policies and significantly accelerates convergence toward profitable operational strategies.

The third layer consists of a Sequence-to-Sequence LSTM-based forecasting model designed to address the partial observability of the operating environment. In practical power-system applications, decision makers do not have access to all variables influencing future system evolution, such as the actions of other market participants, hidden system states, or future imbalance conditions. The proposed encoder-decoder LSTM architecture processes historical observations and learns temporal dependencies in the system dynamics, generating forecasts of future imbalance volumes and prices. Beyond its forecasting capability, the model serves as a state-enhancement mechanism by providing the decision-making framework with information about latent or unobserved system behaviour. In this way, the supervised-learning component helps reconstruct the underlying system state and supplies richer contextual information than would be available from current observations alone. The resulting augmented state representation enables the autonomous decision-making process to anticipate future developments and operate more effectively in partially observable environments.

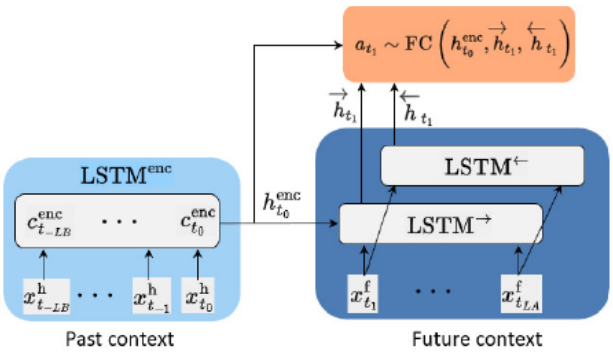

Figure - Sequence-to-sequence LSTM-based model used to infer unobserved dynamics of the partially observable environment.

The resulting methodology can therefore be viewed as a complete learning ecosystem: supervised learning provides information and prior knowledge, while reinforcement learning transforms this information into operational decisions.

3. Main results and perspective

The results demonstrate that combining supervised learning and reinforcement learning leads to substantial improvements in both learning efficiency and operational performance. Forecast information improves the agent's ability to anticipate future conditions, while expert-guided trajectories significantly accelerate training convergence. The resulting hybrid framework learns more rapidly and achieves higher profitability than purely reinforcement-learning-based alternatives.

The study further demonstrates that combining forecasting and reinforcement learning provides a practical mechanism for addressing partial observability. Rather than relying exclusively on historical observations, the agent can exploit predictive information about future system evolution. This significantly improves decision quality in environments characterised by uncertainty and hidden dynamics.

Table - Mean daily profits and standard deviations σp∗ of actions p∗ over the test set, for different forecast noises, averaged over 10 different perturbed realizations per agent.

σ_ϵ (%)		*SafeSAC*	*SACrew*	*SAC*	*MPC*
10	Profit (€/day)	8,109	7,271	5,891	4,558
	σ_{p^*} (MW)	*11.4*	*12.2*	*13.4*	*16.2*
50	Profit (€/day)	6,598	6,432	4,912	3,344
	σ_{p^*} (MW)	*10.2*	*10.2*	*10.6*	*16*
100	Profit (€/day)	5,008	5,029	4,117	1,863
	σ_{p^*} (MW)	*7.6*	*7.2*	*8.4*	*16.2*

From the perspective of DISCRETE, the deliverable provides an important bridge between forecasting, optimisation, and autonomous control. The framework illustrates how supervised learning can provide the situational awareness required by reinforcement-learning agents while RL provides the sequential decision-making capability that forecasting methods alone cannot deliver.

Overall, D2.4 contributes to the project by demonstrating how supervised learners and reinforcement-learning agents can be integrated within a unified framework for real-time power-system operation. The deliverable establishes a methodological pathway toward intelligent decision-support systems that combine forecasting, learning, optimisation, and control within a single data-driven architecture. Future developments may extend this framework toward uncertainty-aware RL, risk-sensitive control, multi-agent coordination, and large-scale transmission-system applications, directly supporting the DISCRETE vision of secure, adaptive, and autonomous operation of future electricity networks.