Transparent Uncertainty Quantification for Offline Reinforcement Learning in Chlor-Alkali Control

Jesse Thibodeau, Étienne Tétreault-Pinard, Said Berriah
{jesse.thibodeau, etienne.tetreault-pinard, said.berriah}@r2.ca

Recherche 2000, Inc.

Abstract

In industrial control settings, practitioners need not only accurate off-policy evaluation (OPE) but also transparent, target-unit-based uncertainty estimates. We introduce VALOR (Validation via Linear Offline Residuals), a lightweight protocol that fits linear surrogate models to production data and uses Mahalanobis-based residual sampling to generate confidence intervals in physical key performance indicator (KPI) units. We apply VALOR to different set-point recommendations made by RL agents trained on a large chlor-alkali plant dataset, and discuss how it enables engineers and plant operators to compare policies with clear, bias-corrected return estimates, facilitating informed, risk-aware decision making, helping bridge the gap between RL research and industrial adoption.

1 Introduction

Chemical manufacturing is the backbone of modern industry, converting bulk feedstocks into the polymers, solvents, fertilizers, and specialized reagents that enable everything from food production to electronics fabrication. Because many of these transformations rely on large-scale electro- or thermochemical steps, the sector is simultaneously a cornerstone of the global economy and one of its most energy-intensive segments. Chlor-alkali electrolysis, defined as the membrane-cell production of chlorine, caustic soda, and hydrogen, underpins $\sim 2\%$ of world chemical sales, yet it is notoriously electricity-hungry. Each tonne of chlorine demands $2.2 - 2.6$ MWh of power, driving a total load of ~ 150 TWh/yr for the aggregate of plants globally. To place this in perspective, industry consumes about 22 PWh of electricity annually, of which 41.9% (≈ 9.2 PWh) is used by manufacturing and heavy processing (International Energy Agency, 2022); chlor-alkali alone therefore accounts for roughly 1.6% of all industrial manufacturing electricity. Even marginal efficiency gains—fractions of a kilowatt-hour per tonne, if adopted industry-wide, translate into multi-million-dollar savings and sizeable CO_2 abatements (Eurochlor, 2023).

Classical control architectures rely on nested PID loops and heuristic set-point tables tuned to nominal operating conditions. These strategies degrade when chemical composition drift, membrane ageing, and volatile power tariffs shift the true optimum away from its calibrated set-point. Reinforcement learning (RL) (Sutton et al., 1998) promises adaptive policies, but three deployment blockers persist in safety-critical plants:

(i) **Transition modeling**: training on existing data logs inhibits exploration and requires offline methods to estimate transition probabilities.

(ii) **Policy opacity**: black-box neural policies erode the confidence of engineers who must audit every deviation from long-established practice.

(iii) **Contextual risk**: managers need statistically relevant error bars before accepting any setpoint recommendation likely to affect hardware component life or product yield.

While each of these challenges is fundamental in real-world RL applications, we specifically address the second and third, as they pertain to policy trustworthiness rather than performance itself. For a high-stakes use case such as chlor-alkali, establishing a common language for model validation is paramount to widespread adoption. Off-policy evaluation (OPE) should therefore be, where possible, grounded in transparent and interpretable statistical methods, allow various domain experts to audit, and ultimately trust, any AI-generated operation policy.

Contributions. This paper introduces VALOR (Validation via Linear Off-policy Residuals), an OPE protocol for validating control recommendations in industrial systems such as chlor-alkali electrolysis, where relationships between control variables and performance objectives can be represented via linear models or linear composites. The method combines interpretable linear surrogates with empirical residual analysis and Monte Carlo sampling to estimate uncertainty-adjusted key performance indicator (KPI) values resulting from a candidate policy. Thus, our contributions are:

(i) A novel validation protocol, VALOR, that produces statistically grounded, bias-corrected confidence intervals for evaluating RL-derived control actions;

(ii) Insights gained from testing VALOR on data from a real chlor-alkali plant data historian;

(iii) A practical argument for favoring transparent, low-complexity uncertainty quantification frameworks in industrial deployments, where auditability and safety remain paramount.

We apply our protocol to a behaviour clone (BC) and a twin-delayed deep deterministic policy gradient agent with BC regularizer (TD3+BC) trained on a full-scale chlor-alkali plant data historian, demonstrating how VALOR enables robust, engineer-auditable decision support. As a caveat, for anonymity, this work provides very limited detail on the specific optimization problem and solution methods, instead focusing analysis on our proposed evaluation protocol.

2 Related Work

2.1 Off-Policy Evaluation

Off-policy evaluation (OPE) seeks to estimate the value of a target policy using a fixed batch of logged data. Three strands dominate the literature:

(i) *Importance-Sampling (IS)* methods provide unbiased estimates but exhibit exponential variance in horizon length (Precup et al., 2000). Doubly-robust and weighted estimators (Jiang & Li, 2016; Farajtabar et al., 2018; Thomas & Brunskill, 2016) reduce variance if accurate models or Q-functions are available.

(ii) *Model-based OPE* fits an approximate dynamics model to generate synthetic rollouts (Kidambi et al., 2020). Bias arises from model mis-specification, prompting pessimistic or ensemble variants (Ghasemipour et al., 2022).

(iii) *Value-function methods* such as Fitted Q Evaluation (FQE) (Le et al., 2019) regress Bellman targets to learn Q^π directly (Uehara et al., 2020). They are data-efficient and stable, but inherit the opacity of deep function approximators and inherently don't provide uncertainty measures.

Benchmark studies (Dann et al., 2014; Fu et al., 2021) show that no single family dominates across tasks: variance control, extrapolation risk, and ease of hyper-parameter tuning remain open challenges.

2.2 Uncertainty Quantification in OPE

Reliable confidence intervals are essential for deploying policies in safety-critical settings. Common techniques include resampling methods such as bootstrap or jackknife on collected trajectories (McIntosh, 2016), which suffer from shrinkage over long horizons; Bayesian value-function approaches (e.g., Bayesian DQN or Bayesian FQI) (Azizzadenesheli et al., 2018), which directly estimate posterior distributions but incur high computational cost; and conformal OPE (Kim et al., 2025), which converts any point-estimate into a finite-sample prediction set but remains rare in continuous-control applications. However, most of these methods express uncertainty in abstract reward units, complicating interpretation by domain experts.

2.3 Industrial Control and Adoption Barriers

Surveys of RL in process industries (Hoi et al., 2021; Nian et al., 2020) attribute slow uptake to three factors: *model opacity*, *lack of hard safety guarantees*, and *mismatch between RL metrics and operational* KPIs. Successful deployments typically rely on either grey-box models augmented with first-principles constraints, or on interpretable surrogates embedded in Model Predictive Control (MPC) loops (Peitz & Dellnitz, 2018). However, a principled framework for *validating* RL recommendations from a transparency angle remains under-explored.

2.4 Positioning of VALOR

Our work bridges the gap between academic OPE and industrial requirements by:

(i) replacing high-capacity Q-networks with highly data-efficient *ordinary-least-squares surrogates* that map control vectors directly to KPI values, retaining transparency;
(ii) introducing a *bias-corrected residual injection* scheme that yields Monte-Carlo confidence intervals in physical KPI units (such as energy use), facilitating engineer review;
(iii) demonstrating compatibility with sequential roll-outs, thereby complementing—rather than replacing—value-based OPE.

To our knowledge, no prior work combines linear surrogates, Mahalanobis-based residual conditioning, and trajectory-level Monte-Carlo propagation into a unified validation protocol, making VALOR a novel contribution to both OPE methodology and industrial RL practice.

3 Background

Chlor-Alkali Electrolysis: A Broad Overview. A modern chlor-alkali plant is, in essence, a large electrochemical factory that turns a continuous stream of purified brine into three commodity products—chlorine

gas, caustic soda (sodium hydroxide), and hydrogen—using massive membrane-cell electrolyzers powered by grid electricity. The site is organized around four unit blocks: (i) *brine preparation*, where raw salt solution is filtered, softened, and chemically conditioned; (ii) the *cell room*, a hall of electrolyzers made up of hundreds of membrane cells operating at tens of kilo-amps each; (iii) *product handling*, where wet chlorine is cooled, dried, compressed, and optionally liquefied, while caustic soda is concentrated by multi-effect evaporation; and (iv) *utilities and balance-of-plant*—cooling water, steam, power-distribution, and process-control systems. Energy costs dominate operating expenditures, so even small efficiency gains in cell-room set-points or utility scheduling translate into substantial economic and environmental benefits. Continuous automation is standard, but most facilities still rely on nested PID loops and fixed rule sheets that must be retuned as membranes age, brine purity drifts, or electricity tariffs fluctuate.

Behaviour Cloning (BC). Behaviour cloning is the simplest offline-RL baseline: it treats policy learning as a supervised-learning task, directly regressing the agent's policy onto the logged behaviour data. Given a batch $\mathcal{D} = \{(s_i, a_i)\}_{i=1}^{N}$ collected under some unknown policy μ, BC fits the parameters θ of a policy π_θ by maximizing the log-likelihood

$$\max_{\theta} \frac{1}{N} \sum_{i=1}^{N} \log \pi_\theta(a_i \mid s_i) \quad \text{(cross-entropy for discrete actions, MSE for continuous).}$$

The resulting policy exactly mimics historical operator actions and is therefore *safe* with respect to distributional shift, but it cannot exceed the performance of the log, nor does it provide value estimates. Consequently, BC serves both as a conservative performance floor and as a source of trajectories whose returns can be estimated by VALOR.

Twin-Delayed Deep Deterministic Policy Gradient with Behaviour-Cloning regularizer (TD3+BC). TD3 (Fujimoto et al., 2018) is an off-policy actor-critic algorithm for continuous control that counters Q-value over-estimation by maintaining *twin* critics and updating the actor less frequently ("twin-delayed"). TD3+BC (Fujimoto & Gu, 2021) adapts TD3 to the offline setting by augmenting the actor loss with a behaviour-cloning term that keeps the learnt policy close to logged actions:

$$\mathcal{L}_{\text{actor}}(\theta) = (1-\alpha)\left[-Q_\phi(s, \pi_\theta(s))\right] \; + \; \alpha \underbrace{\|\pi_\theta(s) - a\|^2}_{\text{BC penalty}}, \qquad (s,a) \sim \mathcal{D},$$

where $Q_\phi(s,\cdot) = \min_{k=1,2} Q_{\phi_k}(s,\cdot)$ is the conservative twin-critic target and $\alpha \in [0,1]$ trades performance improvement against conservatism. For moderate choices $\alpha \approx 0.05$–0.2 the method typically outperforms pure BC while remaining far more stable than unconstrained off-policy RL, so we treat TD3+BC as a state-of-the-art agent for offline continuous-control in this work.

4 Methodology

In this section, we broadly outline our plant control problem on which we learn policies via two solution methods: BC and TD3+BC. Following this, we outline VALOR, which we apply to the learned policies.

4.1 Problem Formulation

Let the plant historian record the full collection of time-stamped variables $\mathcal{S} = \left\{ s_t^{(k)} \,\middle|\, t = 1, \ldots, T,\ k = 1, \ldots, D \right\}$, where D is the total number of available channels and T the horizon length. The control problem

uses only a subset of these channels whose relevance is established by a domain expert,

$$\mathcal{X} = \left\{ x_t^{(j)} \,\middle|\, t = 1, \ldots, T,\ j = 1, \ldots, d \right\} \subseteq \mathcal{S} \quad (d \ll D).$$

Their retained trajectories gives $\mathbf{X} = \left[\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(d)}\right] \in \mathbb{R}^{T \times d}$, where $\mathbf{x}^{(j)} = \left(x_1^{(j)}, \ldots, x_T^{(j)}\right)^\top$. Here $x_t^{(j)}$ denotes the recorded value of the j-th selected variable at time t, and d is the number of variables retained for control. Logged values of $x_t^{(j)}$ reflect the current behavioral policy implemented by operators and determined by a plant's standard operating procedures (SOP).

RL Objective. Let a scalar key-performance indicator (KPI) be defined as a function of the trajectory,

$$\mathrm{KPI}(\mathbf{X}) \;=\; F\big(x_{1:T}^{(1)}, \ldots, x_{1:T}^{(d)}\big),$$

for some plant-specific $F : \mathbb{R}^{T \times d} \to \mathbb{R}$. The control task is to choose inputs so as to find $\max_\pi \; J(\pi) = \mathbb{E}_\pi[\,\mathrm{KPI}(\mathbf{X})\,]$, where the expectation is over trajectories generated by a policy π. A natural KPI could be power consumption under domain constraints.

MDP formalization. For training an RL agent, we model the plant as a Markov Decision Process $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$ with state $s_t = x_t \in \mathbb{R}^d$, action $a_t \in \mathcal{A}$ (set-points), transition kernel P, instantaneous reward $r(s_t, a_t)$ and discount $\gamma \in (0, 1]$. Then $J(\pi) = \mathbb{E}_\pi\big[\sum_{t=1}^{T} \gamma^{t-1} r(s_t, a_t)\big]$ coincides with the KPI when r is chosen appropriately.

Learning approach. Our BC baseline imitates the historical policy π_{hist} (Torabi et al., 2018), while TD3+BC seeks an off-policy π_θ that maximizes $J(\pi)$ (Fujimoto & Gu, 2021), using the same logged dataset. Both methods therefore pursue the KPI-maximization objective defined above. Since RL agent performance itself falls outside this work's focus, we assume these have been trained adequately using default hyperparameters.

4.2 Linear Surrogate Fit

Let f be a linear mapping $f : \mathbb{R}^d \to \mathbb{R}$ modeling the target key performance indicator (KPI), taking the functional form $f = \beta^\top \mathbf{X} + \beta_0$. By carefully selecting a (recent) window of control inputs $\mathbf{X}$ on which to regress KPI, we obtain a simple model of the plant, representative of current operating conditions and equipment wear.

4.3 Validation via Linear Offline Residuals (VaLOR)

Algorithm 1 turns a cheap surrogate model into a statistically-sound *validator* that decides, at each timestamp, whether a candidate set-point vector $\mathbf{x}_{\text{rec}}$ (proposed by BC or TD3+BC) is expected to *improve* the KPI relative to the logged baseline $\mathbf{x}_{\text{base}}$ while providing uncertainty estimates via Monte Carlo sampling of residuals from surrogate predictions on control vectors falling within a ρ-Mahalanobis neighborhood of the recommended control vector.

Algorithm Description. In detail, VALOR first computes KPI estimates $\hat{y}_{\text{base}}$ from behavioral control vector $\mathbf{x}_{\text{rec}}$ (found in plant historian), and $\hat{y}_{\text{rec}}$ from a candidate policy's recommended control vector $\mathbf{x}_{\text{rec}}$ using surrogate model f. Following this, it computes residual r_{base} from $\mathbf{x}_{\text{base}}$ and its corresponding KPI

Algorithm 1 VALIDATION VIA LINEAR OFFLINE RESIDUALS

procedure VALOR($\mathbf{x}_{\text{base}}, \mathbf{x}_{\text{rec}}, f, \rho, N$)
  $\hat{y}_{\text{base}}, \hat{y}_{\text{rec}} \leftarrow f(\mathbf{x}_{\text{base}}), f(\mathbf{x}_{\text{rec}})$ ▷ Surrogate predictions
  $r_{\text{base}} \leftarrow \hat{y}_{\text{base}} - y^{\text{true}}_{\text{base}}$ ▷ Baseline residual
  $B_M \leftarrow \{\mathbf{x}_j : d_M(\mathbf{x}_j, \mathbf{x}_{\text{rec}}) \le \rho\}$; $r_{\text{rec}} \leftarrow \dfrac{1}{|B_M|}\sum_{j \in B_M}(\hat{y}_{\text{rec},j} - y^{\text{true}}_{\text{rec},j})$ ▷ Mahalanobis ball
  Draw $\varepsilon \sim \text{Empirical}\{r_{\text{rec},j}\}_{j \in B_M}$ ▷ Stochastic error
  $\Delta y \leftarrow (\hat{y}_{\text{rec}} - \hat{y}_{\text{base}}) + (y^{\text{true}}_{\text{base}} - \hat{y}_{\text{base}}) + \varepsilon$ ▷ KPI delta
  for each time step **do**
    Bootstrap: repeat Steps 1–6 N times $\rightarrow \{\Delta y^{(n)}\}_1^N$
    CI $\leftarrow$ 95% CI of $\{\Delta y^{(n)}\}$ ▷ Construct 95% CI
    if CI > 0 **then** accept $\mathbf{x}_{\text{rec}}$

value $y^{\text{true}}_{\text{base}}$. This value will be used to capture the surrogate's prediction error at the current operating point, allowing VALOR to subtract that bias when computing the true KPI improvement.

Following this, we construct a Mahalanobis ball consisting of control vectors $\mathbf{x}_j$ falling within a radius ρ of the recommended control vector $\mathbf{x}_{\text{rec}}$, and from them compute a mean residual value r_{rec}. While not strictly necessary, storing r_{rec} is convenient for quantifying the surrogate's expected bias at the recommended point, and can be used to evaluate whether the surrogate's quality deteriorates out-of-distribution. However, in practice, VALOR bootstraps an error estimate ε over N Monte Carlo draws to predict a KPI improvement Δy and construct its corresponding 95% confidence interval CI. Finally, if CI entirely falls above 0 (or some other decision criterion), indicating a KPI improvement, VALOR accepts the policy recommendation.

Bias-variance dial. The Mahalanobis-radius ρ is expected to govern the classical bias-variance trade-off: a smaller ball uses few, very local residuals, while a large ball pools many potentially distant points.

This VALOR layer therefore acts as a lightweight, data-driven safety gate: it certifies each RL recommendation with a transparent confidence interval before any new set-point is deployed.

5 Results and Discussion

Using merely a day of logged control data, we learn linear surrogates for our optimization target, with which we apply VALOR to assess the performance of two different learning agents: a BC and a TD3 agent with BC regularizer. Figure 1 reveals that regressing our KPI on control variables $\mathbf{X}$ obtains an almost-perfect fit, with R^2 just below 1.0 over a 7-day evaluation window. Given that our KPI measure itself is largely grounded in linear expressions established in electro-chemical theory, this high goodness-of-fit is unsurprising. However, equipment deterioration and other confounders may affect the magnitudes of β, therefore a linear regression is nonetheless necessary to obtain a representation of the true current state of the plant. Having trained

	R^2	RMSE	MAE
Linear Regression	0.9999	0.0012	0.0010

Table 1: Normalized KPI regression diagnostics on the evaluation window ($N = 10081$).

models over historical data logs, we test them over the same evaluation timeframe, and visualize policy recommendations corresponding to BC and TD3+BC agents in Figure 1.

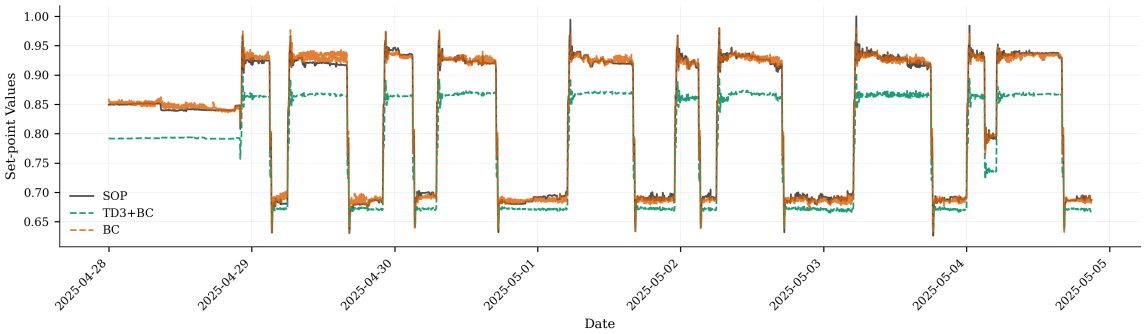
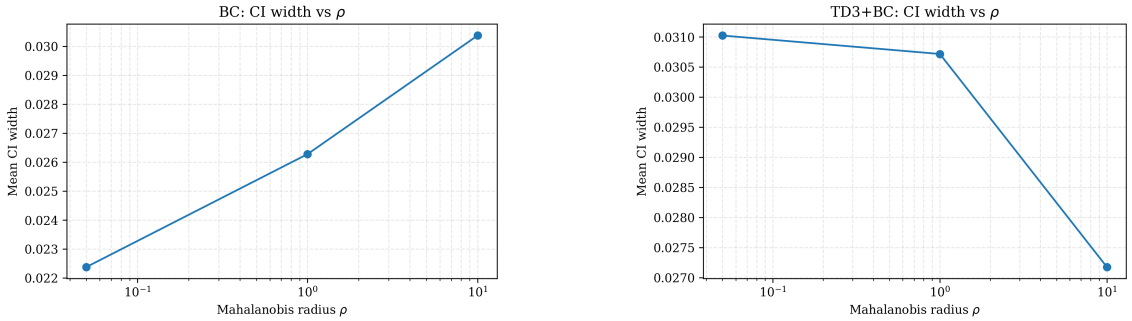


Figure 1: Set-point recommendations for candidate policies for a single control vector (normalized).

5.1 OPE with VALOR

Interpretability & Auditability. Our linear-surrogate design exposes an explicit mapping $\hat{\text{KPI}} = f(\mathbf{X})$ and a closed-form decomposition of bias and stochastic error. A plant engineer can therefore trace every numerical contribution in Algorithm 1 to the final confidence interval.



(a) BC

(b) TD3+BC

Figure 2: (a) CI width increases with ρ for the BC agent and (b) decreases for the TD3+BC agent.

CI-width and Mahalanobis Radius ρ. TD3+BC recommendations lie further from the logged data so a small ρ produces very few local residuals and hence large sampling variance; as ρ grows, more neighbors (even if somewhat distant) are included, reducing Monte-Carlo variance and narrowing the CI. Conversely, BC actions remain close to the logged policy, so at small ρ it already draws from many similar residuals (low variance); increasing ρ forces it to include more heterogeneous, farther-away residuals, which raises sampling variance and widens the CI. Figures 2a and 2b report 95 % CI widths on set-point recommendations from both agents using VALOR across 7 held-out production days.

Ease of Stakeholder Communication. For a power consumption task, a simple acceptance criterion could be: "CI below zero implies expected savings". This aligns with existing energy-efficiency dashboards. In

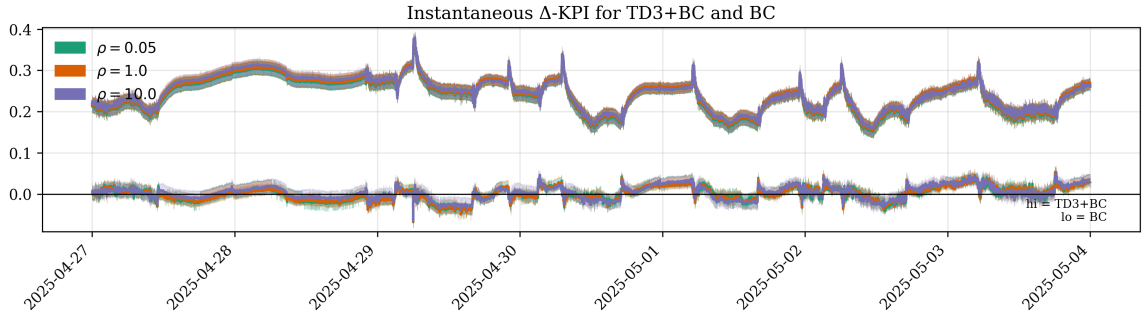


Figure 3: Comparison of estimated KPI improvements from BC and TD3+BC policies.

addition, this produces reward estimates that require no additional mapping to process KPIs, unlike value-based OPE methods like Fitted-Q Evaluation. Examples of estimated KPI improvements for different values of ρ are found in Figure 3.

Sequential Validity. Although VALOR is inherently one-step, repeating the procedure over an evaluation horizon and summing ΔKPI_t yields a trajectory-level estimate, $\Delta\text{KPI}_{1:T} = \sum_{t=1}^{T} \Delta\text{KPI}_t$. This approximation is valid when (i) dynamics are slow or weakly coupled and (ii) control vectors stay within the surrogate's training envelope, which tends to be the case when action-level constraints are imposed.

5.2 Limitations and Future Work

While we find our method to be promising for our use case, VALOR has limitations which should be highlighted. Firstly, while the instanteneous reward estimates are beneficial for interpretability, compounding errors may arise in highly coupled processes. Thus, incorporating a lightweight grey-box transition surrogate is a promising extension. Secondly, current Monte-Carlo draws treat residuals as i.i.d. across timesteps. A block-bootstrap variant could capture temporal correlation. Thirdly, while linearity contributes to transparency, some operating regions may require second-order terms, notably during operating mode transitions. A piecewise-linear or sparse-polynomial surrogate could be be explored with minimal sacrifice in auditability. Finally, A/B testing and collecting qualitative feedback from plant engineers on interpretability of generated risk-assessments is vital for adoption.

6 Conclusion

VALOR contributes to the industrial RL literature by offering a transparent workflow for estimating and auditing policy returns in engineering-relevant units. By combining linear surrogates with a Monte Carlo residual-sampling correction, VALOR produces bias-adjusted confidence intervals that plant engineers and operators can interpret directly in terms of operational KPIs. In our chlor-alkali electrolysis case study with BC and TD3+BC agents, we show that VALOR's clearly communicated uncertainty bounds allow domain experts to compare candidate policies with confidence. Beyond chlor-alkali, this simplicity addresses key barriers to RL deployment in energy-intensive processes, taking steps towards transparent, interpretable and auditable OPE.

References

Kamyar Azizzadenesheli, Emma Brunskill, and Animashree Anandkumar. Efficient exploration through bayesian deep q-networks. In *2018 Information Theory and Applications Workshop (ITA)*, pp. 1–9. IEEE, 2018.

Christoph Dann, Gerhard Neumann, and Jan Peters. Policy evaluation with temporal differences: A survey and comparison. *The Journal of Machine Learning Research*, 15(1):809–883, 2014.

Eurochlor. Electrolysis and production costs: Membrane process competitiveness study. Technical report, Eurochlor, 2023. URL https://www.eurochlor.org/wp-content/uploads/2018/06/12-Electrolysis-production-costs-November-2023.pdf.

Mehrdad Farajtabar, Yinlam Chow, and Mohammad Ghavamzadeh. More robust doubly robust off-policy evaluation. In *International Conference on Machine Learning*, pp. 1447–1456. PMLR, 2018.

Justin Fu, Mohammad Norouzi, Ofir Nachum, George Tucker, Ziyu Wang, Alexander Novikov, Mengjiao Yang, Michael R Zhang, Yutian Chen, Aviral Kumar, et al. Benchmarks for deep off-policy evaluation. *arXiv preprint arXiv:2103.16596*, 2021.

Scott Fujimoto and Shixiang Shane Gu. A minimalist approach to offline reinforcement learning. *Advances in neural information processing systems*, 34:20132–20145, 2021.

Scott Fujimoto, Herke Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In *International conference on machine learning*, pp. 1587–1596. PMLR, 2018.

Kamyar Ghasemipour, Shixiang Shane Gu, and Ofir Nachum. Why so pessimistic? estimating uncertainties for offline rl through ensembles, and why their independence matters. *Advances in Neural Information Processing Systems*, 35:18267–18281, 2022.

Steven CH Hoi, Doyen Sahoo, Jing Lu, and Peilin Zhao. Online learning: A comprehensive survey. *Neurocomputing*, 459:249–289, 2021.

International Energy Agency. World energy balances 2022. https://www.iea.org/data-and-statistics/data-product/world-energy-balances, 2022.

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *International conference on machine learning*, pp. 652–661. PMLR, 2016.

Rahul Kidambi, Aravind Rajeswaran, Praneeth Netrapalli, and Thorsten Joachims. Morel: Model-based offline reinforcement learning. *Advances in neural information processing systems*, 33:21810–21823, 2020.

Dongwon Kim, Matteo Zecchin, Sangwoo Park, Joonhyuk Kang, and Osvaldo Simeone. Robust bayesian optimization via localized online conformal prediction. *IEEE Transactions on Signal Processing*, 2025.

Hoang Le, Cameron Voloshin, and Yisong Yue. Batch policy learning under constraints. In *International Conference on Machine Learning*, pp. 3703–3712. PMLR, 2019.

Avery McIntosh. The jackknife estimation method. *arXiv preprint arXiv:1606.00497*, 2016.

Rui Nian, Jinfeng Liu, and Biao Huang. A review on reinforcement learning: Introduction and applications in industrial process control. *Computers & Chemical Engineering*, 139:106886, 2020.

Sebastian Peitz and Michael Dellnitz. A survey of recent trends in multiobjective optimal control—surrogate models, feedback control and objective reduction. *Mathematical and computational applications*, 23(2): 30, 2018.

Doina Precup, Richard S Sutton, and Satinder Singh. Eligibility traces for off-policy policy evaluation. In *ICML*, volume 2000, pp. 759–766. Citeseer, 2000.

Richard S Sutton, Andrew G Barto, et al. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

Philip Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *International conference on machine learning*, pp. 2139–2148. PMLR, 2016.

Faraz Torabi, Garrett Warnell, and Peter Stone. Behavioral cloning from observation. *arXiv preprint arXiv:1805.01954*, 2018.

Masatoshi Uehara, Jiawei Huang, and Nan Jiang. Minimax weight and q-function learning for off-policy evaluation. In *International Conference on Machine Learning*, pp. 9659–9668. PMLR, 2020.