energies

MDPI

Article

Hierarchical Multi-Objective Dispatch of a Multi-Use Floating Offshore Wind–Hydrogen–Aquaculture System with OpenFAST-Informed Environmental Modeling

Mao Lyu [1,2,*], Chunxiang Fu [1,2], Xiong Zhou [3], Huaijie He [1,2], Lingang Yang [1] and Xiaohe Wang [1]

[1] PowerChina Huadong Engineering Corporation Limited, Hangzhou 311122, China; fu_cx@hdec.com (C.F.); hehuaijie2016@163.com (H.H.); yang_lg@hdec.com (L.Y.); wang_xh8@hdec.com (X.W.)
[2] PowerChina Jiangsu Engineering Corporation Limited, Nanjing 210098, China
[3] CTG Jiangsu Energy Investment Co., Ltd., Nanjing 210098, China; zhouxiong2026@163.com
* Correspondence: lv_m@hdec.com

Abstract

Co-locating hydrogen production and aquaculture on floating offshore wind platforms requires dispatch policies that jointly optimize energy trading, electrolyzer longevity, and cage water quality—objectives that scalarized reinforcement learning struggles to reconcile. We develop a conservation-closed reduced-order model of a 20 MW-class floating wind system, driven by ERA5 North Sea reanalysis and cross-checked against OpenFAST, and cast dispatch as a Markov decision process. Scalarized twin delayed deep deterministic policy gradient (TD3) agents achieved nominally higher revenue than rule-based control (+0.033 M€; Hedges' $g = 0.94$) but converged to a degenerate, hydrogen-abandoning equilibrium in 7% of independent runs (2/27). A hierarchical TD3 agent (H-TD3) with an annealed logarithmic Nash-bargaining reward showed no collapse across 20 seeds (Fisher's exact test, $p = 0.32$), reached 95% of its final physical-objective level at the earliest observable episode in every seed, and exhibited an order-of-magnitude smaller plateau dispersion than scalarized baselines, with revenue statistically indistinguishable from rule-based control (1.216 ± 0.040 M€). Over 1000 h, H-TD3 increased cage biomass by 22.6% and yielded an electrolyzer-subsystem hydrogen cost of €3.05/kg (excluding electricity, wind-farm, and aquaculture costs), or €6.45/kg at €60/MWh wind electricity. These results frame multi-objective offshore dispatch as a risk-control problem: Nash-bargaining rewards produced fewer collapses than linear scalarization, though the difference was not significant.

Keywords: floating offshore wind turbine (FOWT); wind-powered hydrogen; ocean multi-use; PEM electrolyzer; aquaculture; Nash bargaining; deep reinforcement learning

Check for updates

Academic Editor: Tonio Sant

Received: 28 August 2026
Revised: 23 October 2026
Accepted: 29 October 2026
Published: 16 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Floating offshore wind turbines (FOWTs) in the 20 MW class are emerging as the reference unit for next-generation deep-water deployment [1,2], where long export-cable distances make in situ energy conversion—power-to-hydrogen [3,4] and co-located aquaculture [5–7]—economically attractive. However, multi-use platforms introduce a dispatch problem absent from single-purpose systems: a 15 MW proton exchange membrane (PEM) electrolyzer, a day-ahead market price process, an aging membrane stack, and a living biomass compartment all interact through the same limited wind resource on time scales ranging from minutes to weeks.

Motion-induced performance degradation of water electrolysis has been demonstrated experimentally at cell and laboratory-platform scales. These experiments report voltage

deviations of up to 10.2% under six-degree-of-freedom marine sloshing for alkaline cells [8] and orientation-dependent losses of up to 23% at low current density for PEM cells [9]; complementary studies characterize two-phase flow under reorientation [10] and vibration [11,12], bubble dynamics in inclined channels [13], and microgravity effects [14]. A 240 h open-ocean sea trial demonstrated direct seawater electrolysis at 5 kWh Nm^{-3} [15], albeit via a phase-transition mechanism that does not transfer to pressurized PEM systems. Platform- and component-level simulations [16,17] capture structural response but do not resolve electrochemical dynamics. Because validated platform-scale data linking motion to PEM stack performance and degradation remain unavailable, this pathway is not modeled here: the motion-induced de-rating factor is set to unity (Section 2.2), and platform pitch is carried only as a one-way diagnostic proxy that does not alter electrolyzer performance, degradation, or dispatch decisions. Embedding experimentally validated motion effects in closed-loop dispatch therefore remains an open problem, which we leave for future work (Section 7).

The dispatch problem is also algorithmically challenging. Deep reinforcement learning (DRL) is a promising approach to stochastic, nonlinear multi-objective control [18,19], yet our experiments show that scalarizing conflicting objectives into a fixed weighted sum can be fragile: in 2 of 27 independently seeded scalarized runs, agents converged to a degenerate equilibrium in which hydrogen production is abandoned entirely. Nash bargaining theory [20]—which steeply penalizes the sacrifice of any single objective—has been applied to energy-sharing problems [21,22]; however, a reproducible, statistically rigorous characterization of its behavior in floating wind systems is still lacking.

This study makes three contributions. First, we present a conservation-closed simulation pipeline for a 20 MW FOWT–15 MW PEM–aquaculture system, with per-step conservation residuals below 10^{-6} and hydrogen production verified against the thermodynamic limit. Second, we empirically characterize the reward–risk trade-off between scalarized and Nash-bargaining objectives—at 9–20 seeds per configuration—and quantify it with exact confidence intervals (CIs), effect sizes, and collapse-rate statistics. Third, we specify a reproducible hierarchical dispatch framework (H-TD3), comprising a complete Markov decision process, a two-level action structure with slew-rate limiting, logarithmic safety barriers, and a pre-specified statistical protocol that includes Holm–Bonferroni correction and Fisher's exact tests for collapse rates. We therefore position the novelty of this work as hierarchical multi-objective dispatch with OpenFAST-informed environmental modeling; platform motion does not enter the electrolyzer, degradation, or dispatch models.

2. System Modeling and Methodology

2.1. System Configuration and Modeling Framework

The system comprises a single 20 MW semi-submersible FOWT, a 15 MW PEM electrolysis plant (3 × 5 MW stacks) with buffer hydrogen storage and a submerged 8000 m^3 Atlantic salmon cage heated using waste heat recovered from the electrolyzer; Figure 1 shows the architecture and the symbiotic energy and mass flows. All models below are physics-based reduced-order representations constructed to be strictly conservation-closed at the 300 s control step; they do not resolve Navier–Stokes-scale fluid dynamics, and the provenance of every parameter is documented in Tables S1 and S2 of the Supplementary Material.

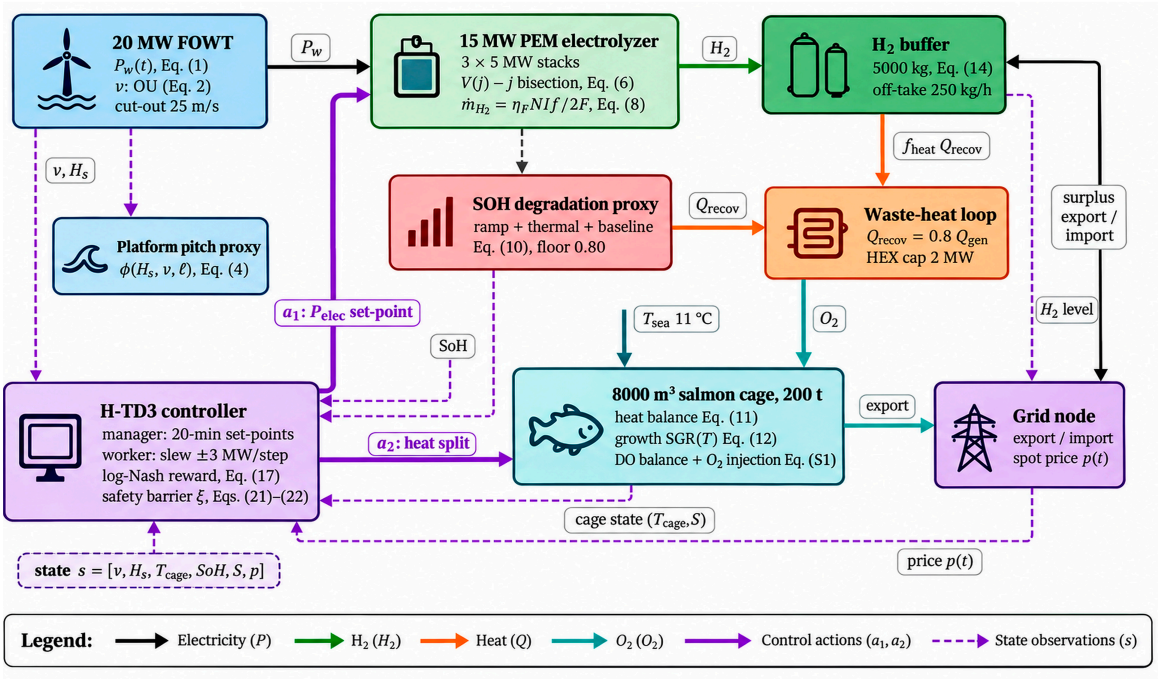


Figure 1. System architecture of the FOWT–hydrogen–aquaculture (FOW-H2-AS) platform. Solid arrows represent energy and mass flows (black = electricity, green = hydrogen, orange = recovered heat, teal = oxygen). Solid purple arrows indicate the control actions issued by the H-TD3 controller (a_1 = electrolyzer power set-point; a_2 = heat allocation fraction to the aquaculture cage), whereas dashed purple arrows represent state observations and feedback to the controller. Every displayed variable corresponds to an implemented model variable described in Section 2. (Prepared with the assistance of ChatGPT (GPT-6; OpenAI, San Francisco, CA, USA).

2.2. Floating Wind Turbine and Met-Ocean Environment

Instantaneous turbine power is represented by an analytic power-curve model:

$$P_w(t) = \begin{cases} 0, & v < v_{ci} \\ 1/2\, \rho_a\, \pi R^2\, C_p(\lambda, \beta)\, \xi_M\, v^3\, \eta_{mech}, & v_{ci} \le v < v_r \\ P_{rated}, & v_r \le v < v_{co} \\ 0, & v \ge v_{co} \quad (\textit{cut-out shutdown}) \end{cases} \tag{1}$$

where $R = 140$ m is the rotor radius (diameter 280 m), $\rho_a = 1.225$ kg/m^3, $\eta_{mech} = 0.95$ is the drivetrain-and-conversion efficiency (typical range 0.93–0.96 [1]), and C_p is capped at $C_{p,max} = 0.482$, calibrated so that the turbine delivers exactly 20 MW at $v_r = 10.5$ m/s. The cut-in and cut-out wind speeds are 3 m/s and 25 m/s, respectively. The motion-induced de-rating factor $\xi_M \in (0, 1]$ is set to unity, which is justified by the aero-hydro-servo-elastic verification campaign reported in Section 5.2, which shows sea-state-driven mean-power spread below 0.42% across all wind speeds; residual second-order effects are acknowledged in Section 7.

Wind speed follows an Ornstein–Uhlenbeck (OU) process, an engineering approximation to the low-frequency content of the Kaimal spectrum at the 5-min control step (full-spectrum verification in Section 4.2):

$$\begin{cases} dv = \theta_v(\overline{v} - v)\, dt + \sigma_v\, dW_v \\ \overline{v} = 10.5 \text{ m/s} \\ \sigma_v^{steady} = \overline{v} \cdot TI \\ TI = 12\% \\ \theta_v^{-1} \approx 30 \text{ min} \end{cases} \tag{2}$$

The parameters are anchored to ERA5 reanalysis [23] over 2014–2023 at three North Sea sites bracketing the study region (57.5° N/1.3° E, 59.5° N/2.5° E, 61.4° N/3.4° E; 87,648 hourly records per site): power-law extrapolation to 150 m ($\alpha = 0.11$) gives annual-mean wind speeds of 10.52–10.64 m/s, consistent with $\overline{v}$ to within 1%; the turbulence intensity (TI) of 12% corresponds to the mid-range of the IEC 61400-1 Class B band [24]; and $-1^{v}_{\theta} = 30$ min is the standard value for low-frequency Kaimal content.

Significant wave height follows a slow OU process tracking a wind-dependent equilibrium (a JONSWAP-magnitude simplification, not a spectral wave model):

$$\begin{cases} dH_s = \theta_H\left(H_s^{eq} - H_s\right) dt + \sigma_H\, dW_H \\ H_s^{eq} = 1.0 + 0.14\, v \\ \theta_H^{-1} \approx 6\,\text{h} \end{cases} \tag{3}$$

Platform pitch response is represented by a linearized proxy rather than a mechanistic hydrodynamic model:

$$\varphi_{rms}(t) = K_{Hs}\, H_s\, (1 - K_L\, \ell(t)) + K_v\, \frac{v}{v_r} \tag{4}$$

where $\ell(t) \in [0, 1]$ is the electrolyzer load fraction, $K_{Hs} = 2.2$ °/m, $K_L = 0.15$, $K_v = 0.4°$ (Table S1), and $v_r = 10.5$ m/s is the rated wind speed. The load term encodes the literature-consistent reduction in platform response under heavier rotor loading [17]; the coupling is one-way and linear. Of the three coefficients, K_{Hs} is the calibration anchor: it is set so that the proxy yields a pitch RMS of ≈5° at the reference sea state ($H_s = 2.5$ m) under full load—a deliberate ≈1.3× margin over the worst-case envelope of the OpenFAST (v4.2.1) campaign quantified below—rather than being regressed from response data. $K_L = 0.15$ fixes the ≈15% full-load reduction in wave response noted above, and $K_v = 0.4°$ contributes a weak steady-trim term at sub-rated wind speeds. Proxy magnitudes are set with reference to the OpenFAST campaign of Section 5.2 as an intentionally conservative upper-bound index, not as a fitted predictor: across 90 load cases, platform-pitch standard deviation spans 0.20–0.92° for operational sea states (P10–P90 of the ERA5 wave climate), with steady trim ranging from −1.0° to +3.7°, giving a campaign pitch-RMS envelope of 0.6–3.8°. The proxy operates at 5.0–5.2° RMS over the 1000 h evaluations, i.e., about 1.3–1.4° (33–37%) above the upper edge of this envelope. The proxy is therefore conservative but is not claimed to be a validated predictor of platform pitch; a point-by-point comparison at the OpenFAST P50 conditions is given in Supplementary Section S6. Pitch is not part of the state, reward, constraints, electrolyzer, or degradation models, so this bias cannot influence any dispatch, hydrogen, revenue, biomass, or economic result; it affects only the diagnostic pitch-RMS values reported in the subsequent results. The proxy is accepted under two explicit criteria: its 1000 h output must not understate the verified OpenFAST pitch-RMS envelope, and it must remain a one-way diagnostic. Limitations of the one-way linear coupling are discussed in Section 7.

2.3. PEM Electrolyzer with Degradation Proxy

The plant comprises three 5 MW-class stacks ($N_{stack} = 3$), each with $n_{cell} = 660$ equivalent series-connected cells of $A_{cell} = 2200$ cm^2, i.e., $N_{cells} = N_{stack} \times n_{cell} = 1980$ cells plant-wide; the three stacks share the load equally, and all plant-level equations below (Equations (7)–(9) and the balances of Section 5.1) use N_{cells}, so that power, hydrogen production, and heat generation refer to the same 15 MW system boundary. The parameters are calibrated so that the system-level higher-heating-value (HHV) efficiency at the 15 MW AC rated point is 68.0% (DC power 13.4 MW after rectifier and balance-of-plant losses, $j \approx 1.60$ A/cm^2, $V_{cell} \approx 1.93$ V; $j_{nom} = 1.7$ A/cm^2 is the nominal stack rating of ≈4.9 MW DC per stack used

for normalization)—which fixes the thermodynamic production ceiling at 259 t per 1000 h of continuous rated operation. The reversible voltage is

$$V_{rev}(T) = 1.229 - 0.85 \times 10^{-3} \, (T - 298.15) \; [\mathrm{V}] \tag{5}$$

and the cell polarization curve is

$$\begin{cases} V_{cell}(j, T, SoH) = V_{rev}(T) + \frac{R_g T}{2F} \ln \frac{j}{j_0} + r_{eff} j + B \ln \frac{j_{lim}}{j_{lim} - j} \\ r_{eff} = r(1 + 0.5\,(1 - SoH)) \end{cases} \tag{6}$$

with parameters from the PEM literature (Table S2): $j_0 = 10^{-7}$ A/cm^2, $r = 0.28\ \Omega{\cdot}$cm^2, $j_{lim} = 2.5$ A/cm^2, $B = 0.05$ V, $T = 80$ °C nominal [4]. Given an AC set-point, the DC power is obtained in closed form from the balance-drive chain.

$$P_{dc} = \frac{(P_{ac} - P_{aux,f})\,\eta_{rect}}{1 + k_{aux}\,\eta_{rect}} \tag{7}$$

(η_{rect}; fixed auxiliary load $P_{aux,f} = 0.70$ MW plus proportional coefficient $k_{aux} = 0.022$, ≈4.7% of rated—within the 3–7% balance-of-plant range [4]), and the current density is then solved by bisection from $P_{dc} = N_{cells} A_{cell}\, j\, V_{cell}(j,T,SoH)$, which is strictly monotone for $j < j_{lim}$. Hydrogen production follows Faraday's law with the same j:

$$\begin{cases} \dot{m}_{H_2} = \eta_F(j)\, N_{cells}\, I\, \frac{M_{H_2}}{2F} \\ I = j\, A_{cell} \\ \eta_F(j) = \mathrm{clip}\left(1 - \frac{j_{cross}}{j},\ 0.90,\ 0.99\right) \end{cases} \tag{8}$$

with crossover current density $j_{cross} = 0.01$ A/cm^2. Hydrogen production (Equation (8)) and cell voltage are computed from the same current solution; here $N_{cells} = 1980$ is the plant-wide cell count, so Equation (8) returns the plant production rate (0.0719 kg/s, i.e., 259 t per 1000 h, at the 15 MW AC rated point, which reproduces the thermodynamic ceiling above), whereas a single 660-cell stack would give one-third of this value; the DC energy balance is verified numerically in Section 5.1. Stack thermal dynamics are first-order:

$$\begin{cases} c_{th}\, \dot{T} = Q_{gen} - Q_{cool} \\ Q_{gen} = N_{cells}\, I\, (V_{cell} - V_{rev}) \\ c_{th} = 1.5 \times 10^8 \ \mathrm{J/K} \end{cases} \tag{9}$$

with a 6 MW cooler ($\tau = 120$ s) and forced de-rating above 95 °C; 80% of Q_{gen} is recoverable to the cage heat loop. Below 10% load, the stacks switch to thermal standby (0.05 MW/stack). The thermal model is lumped at plant level: Q_{gen} uses the plant-wide N_{cells}, and c_{th} and the 6 MW cooler refer to the whole 15 MW plant. The recoverable-heat reference and its sensitivity to this choice are quantified in the following paragraph.

Heat-generation reference and recoverable-heat sensitivity. In Equation (9), Q_{gen} is the total overpotential heat generated inside the stacks and is referenced to the reversible voltage V_{rev}; this Gibbs-referenced definition is the standard electrochemical heat-generation term and, together with Equations (7) and (8), closes the DC allocation identity $E_{dc} = E_{chem} + Q_{gen}$ that is verified numerically in Section 5.1. The net heat duty available for recovery is, however, conventionally referenced to the thermoneutral voltage, $Q_{net} = N_{cells} \times I \times (V_{cell} - V_{tn})$, where $V_{tn}(\mathrm{T}) = 1.481 - 0.88 \times 10^{-3}(\mathrm{T} - 298.15)$ V is the HHV-basis voltage of Table S2; the difference $(V_{tn} - V_{rev})\cdot$I is the reversible reaction heat absorbed into the product enthalpy, which is unavailable to the cooling loop as waste heat unless the product-stream condensation heat is recovered downstream. At the rated point (13.44 MW DC, $j = 1.60$ A/cm^2, $V_{cell} = 1.93$ V, 80 °C), the two references yield $Q_{gen} = 5.22$ MW and $Q_{net} = 3.48$ MW, and the corresponding 80% recovery streams (4.18 and 2.78 MW) both exceed the 2 MW cage heat-exchanger

capacity, so at high load, the heat actually delivered to the cage is exchanger-limited and independent of the reference voltage. A counterfactual replay of all 40 released evaluation runs—20 H-TD3 seeds with the recorded heat-allocation actions held fixed (the conservative static case) and 20 RBC seeds with the hysteresis control law responding to the replayed cage temperature (the exact closed-loop case)—re-integrates the cage thermal balance and growth model with Q_{net} in place of Q_{gen}: the RBC biomass gain is unchanged (28.25% → 28.25%), and the H-TD3 biomass gain shifts from 22.58 ± 4.72% to 23.55 ± 3.91% because the reduced heat input moves the mean cage temperature closer to the 14 °C growth optimum (14.3 → 13.4 °C). The replayed stack temperature never exceeds 80.5 °C, far below the 95 °C de-rating threshold, so the electrolyzer power trajectories are unaffected (summary statistics are provided in Table S3 of the Supplementary Material). The recoverable-heat approximation therefore does not materially affect the aquaculture heating and biomass results; the reported biomass gain of +22.6% is the conservative value under the more generous V_{rev}-referenced recovery.

Stack health is represented by a reduced-order degradation proxy [25–30]:

$$\begin{cases} \dot{SoH} = -\left(a\left|\dot{I}/j_{nom}|_{norm} + b\,\exp\,\frac{T-T_{crit}}{\tau_T} + c\right.\right) \\ a = 4\times10^{-9} \\ b = 1\times10^{-8} \\ c = 2.2\times10^{-9}\ \mathrm{s}^{-1} \end{cases} \tag{10}$$

The baseline term c is set at the mid-range of published PEM decay rates (0.1–2.5%/1000 h); the state of health (SoH) falls from 1.0 to the replacement threshold 0.80 in approximately 15,640 equivalent full-load hours, which feeds the stack-replacement amortization in Section 6. The restriction to ramp, thermal, and baseline terms follows the dynamic PEM degradation literature [25–30]; unresolved mechanisms are acknowledged in Section 7. The resulting electrolyzer characteristics (polarization curve, Faradaic efficiency, and system efficiency map) are shown in Figure S1 of the Supplementary Material.

2.4. Aquaculture Cage

The cage is modeled as a fully mixed lumped thermal mass (ρ = 1025 kg/m^3, c_p = 3990 J/(kg·K); $\rho V c_p \approx 3.27\times10^{10}$ J/K):

$$\begin{cases} \rho_w V_{cage} c_{p,w}\,\dot{T}_{cage} = Q_{ex} + Q_{solar} - Q_{loss} \\ Q_{loss} = k_{loss}\,(T_{cage} - T_{sea}) \end{cases} \tag{11}$$

where $Q_{ex} = \min(f_{heat}\,Q_{recov},\ Q_{hex}^{\max})$ is the heat delivered through the exchange loop ($f_{heat}\in[0,1]$ is the dispatch action; $Q_{hex}^{\max}$ = 2 MW), Q_{solar} is a diurnal sinusoidal surface flux (peak 150 kW × 0.7 cloud/wave attenuation), and $k_{loss} = 3\times10^5$ W/K represents net exchange through the net-pen and water renewal. Growth follows a temperature-dependent specific growth rate (SGR),

$$\begin{cases} \dot{B} = SGR(T_{cage})\,B \\ SGR(T) = SGR_{\max}\,\exp\left[-(T-14/4.5)^2\right] \\ SGR_{\max} = 0.6\ \%/\mathrm{day} \end{cases} \tag{12}$$

with the 12–16 °C optimum band and magnitude for large Atlantic salmon under satiation feeding taken from the aquaculture literature [31]; feed cost and feeding strategy are not modeled (Section 7). Dissolved oxygen (DO) follows a first-order reaeration–respiration balance with proportional pure oxygen injection below a 7.0 mg/L threshold (Equation (S1); parameters in Supplementary Section S1.3); diurnal respiration rhythms and feeding-pulse oxygen demand are not modeled. The DO state enters the safety barrier of Section 3.6, and the three cage processes are illustrated in Figure S2.

2.5. Energy Balance, Hydrogen Storage, and Market

At every 300 s step, the platform-wide power balance is closed by the export/import cable:

$$P_w(t)\,dt = P_{elec}(t)\,dt + E_{exp} - E_{imp} \tag{13}$$

with export remunerated and import charged at the spot price. Imported electricity is not assumed to be renewable or certified: the wind-supplied and grid-supplied fractions of electrolyzer energy are reported separately (imports are 2–4% of electrolyzer energy in the 1000 h evaluations and about 70% in the storm scenario of Section 5.8), and only the wind-supplied fraction is regarded as renewable hydrogen in this study. Produced hydrogen enters a 5000 kg buffer and is withdrawn at a constant contracted off-take of 250 kg/h; deliveries beyond buffer capacity are curtailed (spilled):

$$S(t) = \mathrm{clip}\left(S(t-dt) + \dot{m}_{H_2}\,dt - \dot{m}_{off}\,dt,\ 0,\ S_{\max}\right) \tag{14}$$

The day-ahead price follows a mean-reverting jump process calibrated to the simulation average €60/MWh (a modeling construct, not a historical series):

$$\begin{cases} dp = \theta_p(\overline{p} - p)\,dt + \sigma_p\,dW_p + p\cdot(J-1)\,dN \\ \overline{p} = 0.5 \\ J = 2.2 \\ P(dN = 1) = 0.004/\mathrm{step} \end{cases} \tag{15}$$

The normalized price is mapped linearly such that 1.0 corresponds to €120/MWh, with a floor of 0 and a cap of 1.0.

3. Control Framework

3.1. Markov Decision Process

The dispatch problem is formulated as a Markov decision process (S, A, P, R, γ). The transition kernel P is defined by the coupled simulator of Section 2: given (s_t, a_t), the wind, wave, and price processes advance stochastically according to Equations (2), (3) and (15), and all physical states advance deterministically through the plant models of Equations (5)–(14) and Equation (S1). The discount factor is $\gamma = 0.99$. The state and action vectors are given in full in Table 1; both use fixed linear scalings and are not min–max rescaled to $[0, 1]$, so normalized inputs are permitted to exceed unity under extreme conditions (up to 1.6 for wind speed and 2.0 for significant wave height under the process bounds specified in Supplementary Section S4; in the storm scenario of Section 5.8, with a 26.5 m/s storm plateau (realized gusts to 30.1 m/s) and a 9.1 m peak H_s, $v/25$ and $H_s/6$ reach about 1.21 and 1.52). In the baseline 1000 h evaluation window, the normalized wind speed and wave height stay below 0.64 and 0.58 (15.9 m/s and 3.45 m), and the normalized cage temperature spans about 0.40–0.99 (12.0–17.9 °C, Section 5.7). The policy operates directly on this six-dimensional normalized state; no learned state compression is used.

The information flow is strictly unidirectional: the Worker receives the Manager's set-point and the slew constraint; no learned parameters exist at the execution level. In the non-hierarchical configurations, the same single set-point is refreshed every step (300 s). Empirically, across the 20 training runs (environment seeds 5000–5019) and the 20 deterministic 1000 h evaluation rollouts (seeds 90,000–90,019), the normalized wind speed and wave height never exceed unity (training maxima 0.68 and 0.71, i.e., 16.9 m/s and 4.3 m; evaluation maxima 0.66 and 0.66), whereas the normalized cage temperature exceeds unity in 0.24% of training states and 3.39% of evaluation states, up to 1.10 (19.0 °C) —confirming that the fixed linear scalings are not min–max bounds and that the policy operates on inputs outside $[0, 1]$ without resampling or clipping. The normalized storage level remains below 0.07 (352 kg of the 5000 kg buffer), and the SoH scaling spans $[0.94, 1.00]$ over evaluation.

Table 1. Complete state and action specification. The hierarchy is a set-and-hold action structure executed by one policy: the supervisory level (Manager) issues set-points that the execution level (Worker) tracks with slew-rate limiting; both levels share the same state observation. Normalization is a fixed linear scaling; normalized values may fall outside [0, 1] (Section 3.1).

Level	Variable	Bounds (Physical)	Normalization	Update Period
State (shared)	Wind speed	[0, 40]	$v/25$	300 s
State	Significant wave height	[0.3, 12]	$H_s/6$	300 s
State	Cage temperature	[2, 20]	$(T_{cage} - 8)/10$	300 s
State	Electrolyzer health	[0.80, 1.0]	$(SoH - 0.80)/0.20$	300 s
State	Hydrogen buffer level	[0, 5000]	$S/5000$	300 s
State	Electricity price	[0, 1]	identity	300 s
Action (Manager)	Electrolyzer set-point	[0, 15], slew ± 3 MW/step	a1 ∈ [0, 1], to 15 MW	300 s (held 1200 s in hierarchical mode)
Action (Manager)	Heat split to cage	[0, 1] (≤2 MW effective)	identity	300 s (held 1200 s in hierarchical mode)
Execution (Worker)	Slew-limited power tracking	per-step ramp ≤ 3 MW (=0.01 MW/s = 0.6 MW/min)	deterministic executor	300 s

3.2. TD3 Agent

We use twin-delayed deep deterministic policy gradient (TD3) [32] with twin critics, clipped double-Q learning, target policy smoothing, and delayed policy updates, implemented in PyTorch (version 2.13.0, CPU-only) and trained on a CPU (full hyperparameters in Table S4 of the Supplementary Material; network architecture 6→64→64→2 with tanh output for the actor, 8→64→64→1 for each critic). Training uses 96 h episodes (1152 steps) with 2000 warm-up random steps and exploration noise decaying linearly from 0.2 to 0.05 over 60% of the episode budget; early stopping is triggered when the best 50-episode rolling mean fails to improve by 0.5% for 100 consecutive episodes (subject to a minimum of 120 episodes). Evaluation is a 1000 h (12,000-step) rollout with the deterministic policy under evaluation-environment seeds disjoint from training seeds.

3.3. Objectives and the Unified Log-Nash Reward

Three normalized per-step objectives are defined:

$$\begin{cases} J_{H_2} = \dfrac{\dot{m}_{del}}{\dot{m}_{ref}} \\ J_{fish} = \dfrac{SGR(T_{cage})}{SGR_{\max}} \\ J_{eco} = \dfrac{\dot{R}_{net}(t)}{R_{ref}} \end{cases} \tag{16}$$

where $\dot{m}_{ref} = P_{elec}^{rated} \cdot 0.68/(HHV_{H_2}) \approx 0.0719$ kg/s is the rated HHV-referenced production rate, and $R_{ref} = 0.30$ €/s is a reference net-revenue rate (hydrogen sales + export − import − oxygenation cost + waste-heat credit). Note that J_{eco} is a step-wise economic reward: it is an instantaneous revenue rate, not a discounted net-present-value objective; discounting enters only the project-level analysis of Section 6.

The objective actually optimized is the logarithm of the Nash-bargaining product. Maximizing the Nash product $\prod_i \left(\tilde{R}_i - \tilde{d}_i\right)$ over the normalized surpluses is equivalent to maximizing the sum of logarithms; the implemented per-step reward is

$$r(k) = \sum_{i=1}^{3} \lambda_i(k) \log\left(\max(J_i - d_i(k),\, \varepsilon)\right) + \frac{1}{\mu} \ln \xi(k), \quad \varepsilon = 10^{-4} \tag{17}$$

where $d_i(k)$ is the (annealed) disagreement point and $\lambda_i(k)$ the curriculum weight of objective i at training episode k. Three design elements interact as follows:

Curriculum weights. For the first 120 episodes the weights are uniform, $\lambda_i = 1/3$; thereafter, they transition toward bargaining weights that place more mass on objectives whose surplus is currently smallest:

$$\begin{cases} \lambda_i(k) = (1 - \alpha_k)1/3 + \alpha_k\, w_i \\ w_i \propto \frac{1}{\overline{S}_i + 0.05} \\ \alpha_k = \min\left(k/120,\, 1\right) \end{cases} \tag{18}$$

where $\overline{S}_i$ is an exponential moving average of the positive surplus $J_i - d_i$ (decay 0.999/step). This makes the log-form self-balancing: an objective lagging behind its disagreement point receives more weight precisely when it is at risk.

Disagreement points and annealing. The disagreement points are the mean normalized objectives of the rule-based control (RBC) baseline, pre-calibrated over three 1000 h baseline runs: $d^{RBC} = (0.953, 0.995, 1.117)$ for $(J_{H_2}, J_{fish}, J_{eco})$. With fixed $d_i = RBC_d^i$, a randomly initialized policy sits in the $log(\varepsilon)$ floor for all three terms from step one—every gradient carries the same constant, and the learning signal is unreachable. The full hierarchical agent (H-TD3) therefore anneals the disagreement point linearly from zero to the RBC baseline over the same 120-episode schedule:

$$d_i(k) = d_i^{RBC} \cdot \min\left(k/120,\, 1\right) \tag{19}$$

This is a standard reward-shaping device; it does not predetermine the result (the fixed-point variant, TD3+Nash, is retained in the ablation as a separate configuration; Section 5.4).

Scale invariance and its implications. Because the objective is a sum of logarithms of dimensionless normalized surpluses, no objective can be driven deeply negative to obtain marginal gains in another: as any $J_i - d_i \to \varepsilon$, its log term plunges to $\log\varepsilon \approx -9.2$ per step, dominating any plausible gain elsewhere. Under a fixed-weight additive reward, in contrast, the shut-down/all-export policy keeps J_{eco} near its reference indefinitely while sacrificing J_{H_2} and J_{fish}—which is precisely the collapse mode observed empirically in Section 5.4. We emphasize what we do *not* claim: this is an empirical robustness property of the reward design demonstrated by ablation, not a proof that the resulting policy is Pareto-optimal or that the game-theoretic axioms of the exact Nash bargaining solution carry over to approximate DRL on a stochastic non-convex problem.

3.4. Linear Scalarized Baseline Reward

The ablation isolates the reward structure, with the objectives J_i as defined in Equation (16). The plain TD3, hierarchical TD3+Hier, and physics-informed TD3+PIRL configurations optimize

$$r_{lin}(k) = \frac{1}{3} \sum_{i=1}^{3} J_i(k) + \frac{1}{\mu} \ln \xi(k) + w_{pirl} \sum_m |f_m(k)| \tag{20}$$

with $w_{pirl} = 0$ except in TD3+PIRL ($w_{pirl} = 1.0$); the full H-TD3 uses $w_{pirl} = 0.1$ in addition to Equation (17). TD3+Nash uses Equation (17) with fixed disagreement points and uniform

weights; TD3+Hier uses Equation (20) with the hierarchical action structure of Table 1; H-TD3 combines Equation (17) with annealing (Equation (19)), curriculum (Equation (18)), the hierarchical action structure, and $w_{pirl} = 0.1$.

3.5. Physics-Informed Residual as a Numerical Integrity Audit

The physics-informed term of the linear reward (Equation (20)) is built from the residuals f_m of the four conservation balances implemented in the simulator: the wind-power balance (Equation (13)), the electrolyzer DC allocation ($E_{dc} = E_{chem} + Q_{gen}$), an identity of the shared-current formulation of Equations (6) and (8)), the cage heat balance (Equation (11)), and the dissolved-oxygen mass balance (Equation (S1)). Its normalized form is given in Equation (S2). Because the simulator is conservation-closed, these residuals remain at machine precision during operation, so the term functions as a numerical integrity audit of the training rollout rather than as a mechanism that steers the policy toward physical feasibility; no Navier–Stokes or bubble-dynamics residual is included. The empirical consequence—TD3+PIRL is bit-identical to plain TD3 in this implementation—is reported in Section 5.9.

3.6. Safety Formulation: Logarithmic Barrier

Safety is enforced through a barrier term on the smallest normalized slack,

$$\begin{cases} \xi(k) = \max\left(\min(\xi_{DO}, \xi_S, \xi_{SoH}), 10^{-6}\right) \\ r_{safe}(k) = \frac{1}{\mu} \ln \xi(k) \\ \mu = 0.02 \end{cases} \tag{21}$$

$$\begin{cases} \xi_{DO} = \text{clip}\left(\frac{DO - DO_{hard}}{\Delta_{DO}}, 0, 1\right) \\ \xi_S = 1 - S/S_{\max} \\ \xi_{SoH} = \text{clip}\left(\frac{SoH - 0.80}{0.02}, 0, 1\right) \end{cases} \tag{22}$$

with $DO_{hard} = 6.0$ mg/L (biological stress line) and $\Delta_{DO} = 0.5$ mg/L. The sign convention is as follows: since r is maximized, the $+\frac{1}{\mu} \ln \xi$ term tends to $-\infty$ as $\xi \to 0^+$ and thus penalizes approach to any constraint boundary; it is equivalent to the usual interior-point barrier for maximization with the domain constraint $\xi \in (0, 1]$—the same domain-constrained logarithm used for the biological slack, $ln(DO - DO_{hard})$, taken through its normalized, clipped form so the argument is strictly positive. The DO slack is scaled so that the barrier activates only below 6.5 mg/L, preventing the (state-unobservable) oxygenation controller's chatter from acting as critic noise. The barrier is a soft constraint; hard feasibility is additionally guaranteed by the executor (slew limits, temperature de-rating, load floor), and barrier-violation statistics are reported in Section 5. The barrier steepness, its DO-activation threshold, and the curriculum/annealing schedule are illustrated in Figure S3.

4. Simulation Setup

4.1. Training and Evaluation Protocol

Six configurations are compared (Table 2): the rule-based baseline RBC (power-following with hysteresis-based cage temperature control at 13.5–14.5 °C, no learning) and five RL variants. TD3 and TD3+Nash are trained with 18 independent seeds each, H-TD3 with 20, TD3+PIRL with 10, and TD3+Hier with 9; the deterministic RBC is evaluated over 20 independent environment seeds. Evaluation seeds are disjoint from training seeds (seed derivation in Supplementary Section S3). The unequal n (9–20 per configuration) is handled natively by the Welch test family and by bootstrap CIs, which do not assume equal sample sizes; collapse-rate comparisons pool the two independent scalarized configurations (TD3 and TD3+Hier; 27 runs; TD3+PIRL is excluded because it reproduces

TD3 bit-for-bit, Section 5.9) against the log-Nash configurations and are reported alongside per-configuration Clopper–Pearson CIs (Section 5.4). Evaluation is a 1000 h deterministic-policy rollout on an independent realization of the same stochastic processes as training; no cross-site generalization is claimed.

Table 2. Configuration matrix.

Configuration	Reward	Hierarchical Action	PIRL Weight	Disagreement Treatment
RBC	— (rule: power-following + hysteresis heating)	—	—	—
TD3	Linear, Equation (20)	No	0	—
TD3+PIRL	Linear, Equation (20)	No	1.0	—
TD3+Hier	Linear, Equation (20)	Yes	0	—
TD3+Nash	Log-Nash, Equation (17)	No	0	Fixed, $d = d^{RBC}$ uniform weights
H-TD3 (full)	Log-Nash, Equation (17)	Yes	0.1	Annealed, Equation (19); curriculum, Equation (18)

4.2. Environment Processes and Data

All stochastic inputs are seed-controlled synthetic realizations: wind speed and significant wave height follow the OU processes of Equations (2) and (3); the electricity price follows the jump process of Equation (15); sea temperature follows a deterministic seasonal–diurnal cycle (11.0 ± 1.2 °C); and hydrogen off-take is constant at 250 kg/h (full specification in Table S5). No measured met-ocean series enters the RL training loop; the ERA5 reanalysis [23] is used instead to (i) anchor the process parameters of Equations (2) and (3) to observed North Sea statistics (Table S6, Figure S4) and (ii) define the sea states of the independent OpenFAST verification of Section 5.2. The environmental window used in all runs is a cold-season high-wind scenario (mean sea temperature 11 °C), under which waste-heat recovery for cage temperature control is relevant.

5. Results

5.1. Conservation Verification

An independent audit run (seed 777, RBC policy, 1000 h) confirms the conservation properties on which the remaining analysis relies. Total wind generation was 17,355 MWh; the electrolyzer consumed 14,161 MWh, 3485 MWh was exported, and 290 MWh was imported; the per-step residual of the balance (Equation (13)) was below 10^{-6} with a cumulative energy-balance residual of 0.0000%. The DC allocation residual ($E_{dc} = E_{chem} + Q_{gen}$: 12,655 = 7819 + 4836 MWh) and the cage heat-balance residual were also below 10^{-6} per step. The hydrogen mass balance closed exactly (produced 246,251 kg = delivered 246,237 + stored 14 + wasted 0). All configuration means satisfy the thermodynamic ceiling: H-TD3 212.0 t, RBC 246.7 t, TD3 214.2 t, TD3+Hier 199.0 t, TD3+Nash 217.5 t, TD3+PIRL 202.7 t, all ≤259 t per 1000 h at the model's 68% HHV rating. These checks, generated automatically by the pipeline for every results release, are the empirical basis for trusting the production and revenue figures below; the audit is visualized in Figure S5.

5.2. Anchoring Against Aero-Hydro-Servo-Elastic Simulation (OpenFAST, 90 Load Cases)

The reduced-order environment of Section 2.1 rests on two assumptions that a power-curve-plus-proxy model cannot verify internally: (i) that sea-state-induced platform motion

does not materially change the mean power available to the dispatch layer (the $\xi_M = 1$ choice), and (ii) that the linearized pitch proxy of Equation (4) gives response magnitudes that bound those of a fully coupled floating system from above. Both were tested against an independent aero-hydro-servo-elastic campaign built on OpenFAST (v4.2.1) [33] with the IEA 15 MW reference turbine on the UMaine VolturnUS-S semi-submersible platform [34,35], the ROSCO v2.9.0 reference controller [36] with a single disclosed modification (Supplementary Section S6), IEC 61400-1 Class 1B normal-turbulence Kaimal fields generated with TurbSim (v4.2.1, x64; National Renewable Energy Laboratory, Golden, CO, USA), and JONSWAP irregular waves conditioned on the ERA5 joint wind–wave statistics. The 90-case matrix spans 12 mean wind speeds from cut-in to cut-out (3–25 m/s), with three sea-state quantile families (P10/P50/P90) in the operational band and three turbulence seeds per condition (Figure S4c; full design in Supplementary Section S6). The 15 MW reference model verifies the structure and sea-state sensitivity of the reduced-order formulation—not the 20 MW numerical values, which follow from the analytic power curve of Equation (1).

Findings. Two results bear directly on the reduced-order model. First, mean power is sea-state-insensitive: at every wind speed, the mean generated power differs by less than 0.42% between the P10, P50, and P90 sea states (largest spread at 5 m/s, where power is a small fraction of rating), and the rated-power band (≥15 m/s, 36 cases) delivers 14.97 MW, on average, 99.8% of the 15 MW rating. Whatever platform motion the sea state induces, its effect on the 5-min-mean power available to the electrolyzer is therefore sub-percent, which directly supports $\xi_M = 1$. Second, the pitch proxy is conservative but not quantitatively accurate: the operational-band pitch standard deviation spans 0.20–0.92° around a steady trim of −1.0° to +3.7° that peaks near rated wind speed (Table S7), and the combined campaign pitch-RMS envelope (0.6–3.8°) lies below the proxy's 1000 h realizations (5.06–5.20° across configurations, Table 3). The proxy exceeds the upper edge of the verified envelope by about 1.3–1.4° (33–37%); at the P50 conditions of Table S7, it overpredicts the OpenFAST pitch standard deviation by a factor of about 3–13 (Supplementary Section S6). It is therefore used as a conservative diagnostic index, not as a validated pitch predictor (acceptance criteria in Section 2.2). At low wind speed, the OpenFAST pitch response is turbulence, rather than wave-dominated (pitch standard deviation of 0.20–0.23° across H_s = 0.64–2.32 m at 3 m/s); the linear proxy does not reproduce this ordering because its H_s term scales linearly with wave height at all wind speeds, which is a further reason to treat it as a conservative index only. Mooring and extreme-response screening results, which do not enter the dispatch model, and the limitations of the campaign are given in Supplementary Section S6. Figure 2 summarizes the campaign.

Table 3. Ablation results, 1000 h deterministic evaluation, mean ± SD over seeds. n = 20 (RBC, H-TD3), 18 (TD3, TD3+Nash), 10 (TD3+PIRL; a diagnostic replicate of TD3, Section 5.9, excluded from pooled collapse inference), 9 (TD3+Hier); unequal n is handled natively by the Welch/bootstrap statistics of Section 5.5.

Metric	RBC (n = 20)	TD3 (n = 18)	TD3+PIRL (n = 10)	TD3+Hier (n = 9)	TD3+Nash (n = 18)	H-TD3 (n = 20)
H_2 produced (t)	246.67 ± 0.77	214.19 ± 55.67	202.66 ± 72.23	198.97 ± 75.36	217.45 ± 65.31	212.03 ± 53.29
— median (range)	246.7 [244.9–247.9]	231.8 [0–241.8]	222.8 [0–240.9]	221.0 [0–244.6]	247.9 [0–258.8]	239.7 [65.5–250.2]
Revenue (M€)	1.206 ± 0.008	1.238 ± 0.049	1.230 ± 0.055	1.224 ± 0.062	1.217 ± 0.042	1.216 ± 0.040
Biomass (t)	256.49 ± 0.00	239.04 ± 7.47	238.46 ± 7.39	243.95 ± 8.73	238.48 ± 8.51	245.16 ± 9.69
Biomass gain vs. 200 t initial (%)	28.2 ± 0.0	19.5 ± 3.7	19.2 ± 3.7	22.0 ± 4.4	19.2 ± 4.3	22.6 ± 4.8

Table 3. *Cont.*

Metric	RBC ($n = 20$)	TD3 ($n = 18$)	TD3+PIRL ($n = 10$)	TD3+Hier ($n = 9$)	TD3+Nash ($n = 18$)	H-TD3 ($n = 20$)
SoH decay (%/1000 h)	1.279 ± 0.000	1.192 ± 0.298	1.129 ± 0.398	1.136 ± 0.426	1.142 ± 0.317	1.279 ± 0.002
Electrolyzer energy (MWh)	14,187 ± 50	12,175 ± 3155	11,520 ± 4062	11,260 ± 4226	12,508 ± 3747	12,009 ± 3179
Export (MWh)	3519 ± 59	5648 ± 3067	6247 ± 3977	6489 ± 4066	5381 ± 3550	5782 ± 2968
Import (MWh)	284 ± 13	396 ± 150	358 ± 158	343 ± 195	462 ± 242	370 ± 239
Pitch RMS (°)	5.06 ± 0.06	5.17 ± 0.15	5.18 ± 0.19	5.20 ± 0.22	5.16 ± 0.23	5.17 ± 0.16
Convergence plateau, $\bar{J}$ (episodes)	—	220 ± 71	201 ± 63	264 ± 72	270 ± 77	212 ± 8.4
Convergence plateau, $\bar{J}$ (episodes)Time to 95% of final $\bar{J}$ (episodes)	—	220 ± 7162 ± 37	201 ± 6365 ± 44	264 ± 7253 ± 8.5	270 ± 7778 ± 55	212 ± 8.450.0 ± 0.0

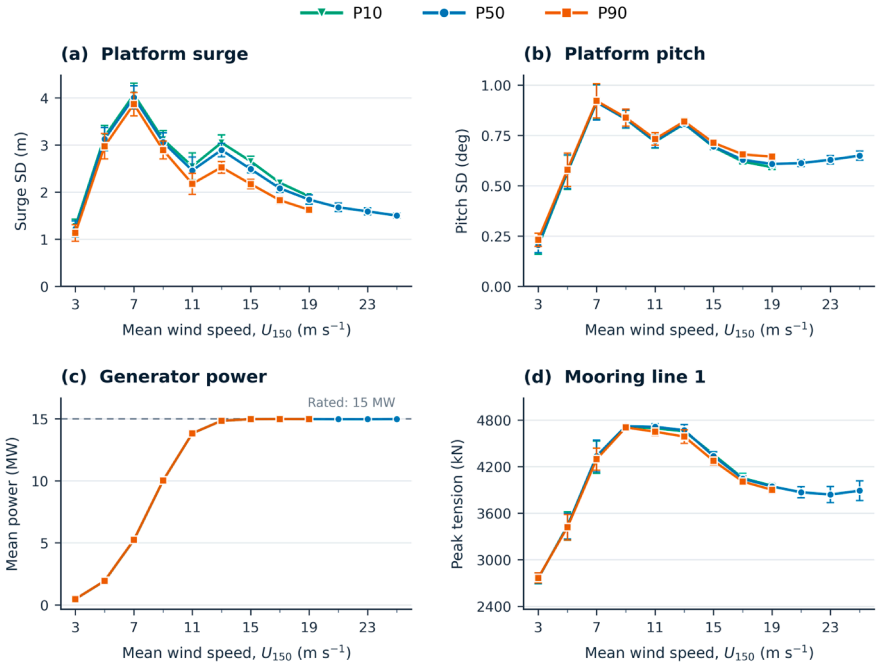


Figure 2. OpenFAST verification campaign (90 load cases, IEA 15 MW reference turbine on the UMaine VolturnUS-S semi-submersible, ROSCO v2.9.0, IEC 61400-1 Class 1B NTM Kaimal turbulence, three seeds per condition). Mean ± SD over turbulence seeds versus mean wind speed for the P10/P50/P90 sea-state families: (**a**) platform surge standard deviation, (**b**) platform pitch standard deviation, (**c**) mean generated power with the 15 MW rated line, and (**d**) peak fairlead tension. The near-coincidence of the three sea-state curves in (**c**) (spread < 0.42%) is the empirical basis for the $\xi_M = 1$ choice in the dispatch model.

5.3. Six-Configuration Ablation

Table 3 reports the 1000 h evaluation metrics as mean ± SD over seeds (± is the sample SD across independent seeds, not a standard error; for the deterministic RBC, the SD reflects environmental stochasticity only).

Three observations follow, none of which implies uniform superiority of any configuration:

Plain TD3 achieved the highest mean revenue of all six configurations (1.238 ± 0.049 M€), directionally above RBC (+0.033 M€; Hedges' $g = 0.94$; bootstrap 95% CI [0.009, 0.053] M€, which excludes zero; Mann–Whitney U (MWU) $p = 4.6 \times 10^{-4}$), although the Holm-corrected Welch test does not confirm the difference ($p = 0.105$; Section 5.5). The upside is paired with exposure to the collapse pathology: one of 18 TD3 seeds converged to the degenerate shut-down equilibrium (Section 5.4).

RBC produced more hydrogen and more biomass than every RL configuration (246.7 t and +28.2% biomass vs. H-TD3's 212.0 t and +22.6%). For H-TD3, both deficits are quantified in Section 5.5: hydrogen −34.6 t (bootstrap CI [−59.1, −13.9] t, excluding zero) and biomass gain −5.7 percentage points (Holm-corrected $p = 0.001$, $g = -1.62$). This is mechanistically expected: RBC keeps the electrolyzer near full load, maximizing both hydrogen throughput and recoverable waste heat, whereas the RL policies trade electrolyzer energy against price-responsive export (H-TD3 exported 5782 ± 2968 MWh vs. RBC's 3519 ± 59 MWh). The learned policies' value proposition is therefore not higher output but a different—and for H-TD3, statistically indistinguishable—revenue mix, plus robustness properties that RBC's SD cannot express because RBC has no policy variance by construction. The biomass ordering and magnitudes are robust to the heat-recovery reference voltage: under the thermoneutral-referenced counterfactual of Section 2.3, RBC remains at +28.2% and H-TD3 shifts to +23.5 ± 3.9%, narrowing the gap from 5.7 to 4.7 percentage points without changing the ordering.

H-TD3 converged most consistently in the physical-objective space. Defining the equal-weight objective $\bar{J} = \left(J_{H2} + J_{fish} + J_{eco}\right)/3$, which has an identical definition across reward designs, every one of the 20 H-TD3 seeds first reached 95% of its final $\bar{J}$ at episode 50—the earliest observable episode given the 50-episode window—with zero seed dispersion, whereas the scalarized configurations averaged 53–78 episodes with seed SDs of 8.5–54.5. The $\bar{J}$-plateau of H-TD3 is statistically indistinguishable in location from TD3 (Mann–Whitney $p = 0.85$) but an order of magnitude more consistent across seeds (SD 8.4 vs. 71.4 episodes; Levene $p = 1 \times 10^{-4}$; Section 5.6).

Per-seed distributions of all six metrics behind Table 3 are given in Figure S6.

A methodological caveat applies to every cross-configuration statement in this paper: raw training/evaluation rewards are not comparable across reward designs (the log-Nash evaluation reward is on a log scale and takes large negative values by construction); all comparisons above and below use physical and economic metrics only.

5.4. Policy Degradation and Collapse Analysis

The central result of the ablation concerns policy collapse. Four of the 75 RL runs (3 of the 65 statistically independent runs, see below) converged to a degenerate shut-down equilibrium with a common signature: zero hydrogen production, electrolyzer energy exactly at the 150 MWh standby floor over 1000 h (3 stacks × 0.05 MW × 1000 h, i.e., the electrolyzer fully off), maximal export with zero import, and essentially zero degradation (SoH decay $\approx 8 \times 10^{-6}$ %/1000 h). The policy is, in effect, "sell all the wind, never run the electrolyzer." Three of these are scalarized-reward runs (one seed each of TD3 (seed 4), TD3+Hier (seed 0), and TD3+PIRL (seed 4)). Because TD3+PIRL seed 4 is the bit-identical counterpart of the collapsed TD3 seed 4 (Section 5.9), i.e., the same training realization, it is not counted as an independent observation: the independent collapse count of the scalarized family is 2 of 27 runs (TD3, $n = 18$; TD3+Hier, $n = 9$), and TD3+PIRL is treated as a diagnostic confirmation of TD3 rather than as a separate group in the pooled statistics. The remaining collapse is 1 of the 18 fixed-disagreement TD3+Nash runs (seed 12), whose equal-weight objective reached 95% of its (degenerate) final level by episode 50, i.e., it converged rapidly onto the degenerate optimum. The H-TD3 agent—the same log-Nash

reward combined with the annealed disagreement schedule and the hierarchical two-level action structure—collapsed in 0 of 20 runs, as did RBC in 0 of 20 (by construction). Collapse also depressed revenue: collapsed scalarized seeds earned about 1.06–1.12 M€ versus about 1.25 M€ for the non-collapsed cluster (Figure S6b; no significance test is reported because only two independent collapsed runs are available) because low-price export cannot compensate for the lost hydrogen revenue. Table 4 lists the collapsed runs and the resulting collapse rates.

Table 4. Collapse incidence by configuration (1000 h evaluation; the number of independent runs varies by configuration). TD3+PIRL runs 0–4 reproduce TD3 bit-for-bit (Section 5.9); the TD3+PIRL row is shown for completeness only and is excluded from the pooled scalarized family, which therefore contains independent runs only (TD3 & TD3+Hier).

Group	Runs	Collapsed	Rate	CP 95% CI
TD3	18	1	0.056	[0.001, 0.273]
TD3+PIRL (replicate of TD3; not pooled)	10	1	0.100	[0.003, 0.445]
TD3+Hier	9	1	0.111	[0.003, 0.482]
Scalarized family, independent runs (TD3 & TD3+Hier)	27	2	0.074	[0.009, 0.243]
TD3+Nash (fixed disagreement)	18	1	0.056	[0.001, 0.273]
Log-Nash family (pooled)	38	1	0.026	[0.001, 0.138]
H-TD3 (annealed, hierarchical)	20	0	0.000	[0.000, 0.168]
RBC (baseline)	20	0	0.000	[0.000, 0.168]

One-sided Fisher exact tests on the pre-specified comparisons, computed on independent runs only (scalarized family = TD3 & TD3+Hier, 2/27): scalarized family vs. H-TD3, $p = 0.325$; scalarized family vs. TD3+Nash, $p = 0.651$; scalarized family vs. RBC, $p = 0.325$; TD3+Nash vs. H-TD3, $p = 0.474$.

The mechanistic reading of the scalarized collapses is consistent with the reward analysis of Section 3.3: under the additive linear reward, the all-export policy keeps J_{eco} near its reference while sacrificing J_{H_2} and J_{fish}, and the resulting expected return remains locally attractive. Under the log-Nash reward, any surplus driven to zero makes the reward plunge toward $\log\varepsilon$ per step, which renders the shut-down equilibrium maximally punitive—yet the single TD3+Nash collapse demonstrates that this protection is probabilistic, not absolute: the log-Nash reward alone does not make the degenerate equilibrium unreachable, and the mechanism by which that seed escaped the steep log-slope punishment is not resolvable from the released training logs (we do not speculate beyond noting its fast convergence onto the degenerate optimum). What the data do show is that the only configuration with zero observed collapse at $n \geq 18$ is the full H-TD3 agent, which combines the log-Nash reward with the annealed disagreement schedule and the hierarchical action structure; whether its 0/20 reflects those design elements or chance cannot be decided at these sample sizes. Nor does the log-Nash family reduce outcome variance: the hydrogen-production variance ratios are near unity (TD3 vs. TD3+Nash 0.73, Brown–Forsythe $p = 0.53$; TD3 vs. H-TD3 1.09, $p = 0.48$).

We are explicit about the statistical status of the collapse result: the rate differences (2/27 vs. 0/20, 1/18 vs. 0/20) are not themselves significant (Fisher $p = 0.32$–0.65), and a designed power analysis shows that distinguishing a 20% from a 0% collapse rate at 80% power requires ≈30 seeds per group (power 0.66 at 20 vs. 20; 0.83 at 30 vs. 30); at the observed 7.4% vs. 0% rates, the achievable power is lower still, and the 0/20 H-TD3 estimate carries a 95% CI upper bound of 16.8%, i.e., the data cannot rule out collapse rates up to ≈17% for H-TD3 either. We therefore present the collapse observation as an empirical regularity with exact CIs (Table 4)—mechanistically explained for the scalarized family,

observed once under the fixed-disagreement log-Nash variant, and absent in 20 H-TD3 runs—rather than as a statistically confirmed effect.

Figure 3 shows the per-seed production values with the collapse criterion and the Clopper–Pearson incidence intervals.

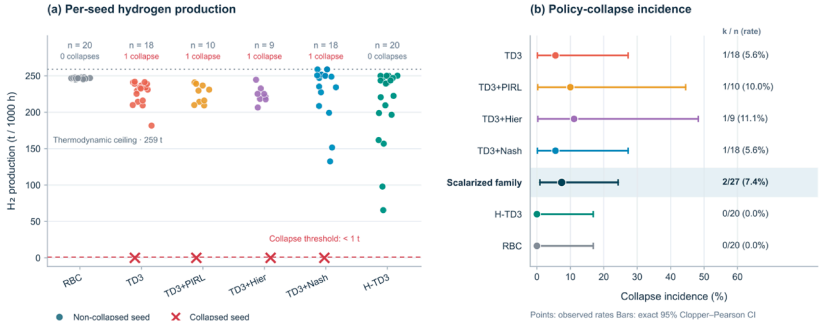


Figure 3. Collapse analysis. (**a**) Per-seed 1000 h hydrogen production by configuration (points = individual seeds; red × = collapsed seeds below the 1 t criterion, dashed line; dotted line = 259 t thermodynamic ceiling); (**b**) collapse incidence with Clopper–Pearson exact 95% CIs per configuration and for the pooled scalarized family (2/27, independent runs only) and H-TD3 (0/20). The single TD3+Nash collapse (seed 12) and the zero-collapse H-TD3 record are both visible; the CIs overlap, which reflects the statistical status of the comparison. TD3+PIRL, a bit-identical replicate of TD3, is excluded from the pooled estimate.

5.5. Statistical Comparisons

Table 5 lists the pre-specified pairwise comparisons on the 1000 h evaluation metrics.

Table 5. Pre-specified pairwise comparisons (1000 h evaluation). Mean difference is (first − second). Welch t-test with Holm–Bonferroni correction across the 11-comparison family; Mann–Whitney U (MWU) as a non-parametric check; bootstrap 95% CI of the mean difference (10,000 resamples, percentile method); Hedges' g effect size. $n = 20$ (RBC, H-TD3) and 18 (TD3, TD3+Nash); unequal n enters natively through Welch statistics and bootstrap resampling.

Contrast	Metric	Mean$_1$ ± SD	Mean$_2$ ± SD	Welch t	p (raw)	p (Holm)	Hedges' g	Bootstrap 95% CI	MWU p
TD3 vs. RBC	Revenue (M€)	1.238 ± 0.049	1.206 ± 0.008	2.81	0.012	0.105	0.94	[0.009, 0.053]	<0.001
	H_2 (t)	214.19 ± 55.67	246.67 ± 0.77	−2.48	0.024	0.193	−0.83	[−61.1, −14.8]	<0.001
H-TD3 vs. RBC	Revenue (M€)	1.216 ± 0.040	1.206 ± 0.008	1.11	0.279	1.000	0.34	[−0.007, 0.028]	0.457
	H_2 (t)	212.03 ± 53.29	246.67 ± 0.77	−2.91	0.009	0.090	−0.90	[−59.1, −13.9]	0.068
	Biomass gain (%)	22.58 ± 4.84	28.25 ± 0.00	−5.23	<0.001	0.001	−1.62	[−7.78, −3.68]	<0.001
TD3+Nash vs. TD3	H_2 (t)	217.45 ± 65.31	214.19 ± 55.67	0.16	0.873	1.000	0.05	[−36.1, 42.5]	0.082
	Revenue (M€)	1.217 ± 0.042	1.238 ± 0.049	−1.44	0.158	0.950	−0.47	[−0.050, 0.008]	0.018
H-TD3 vs. TD3+Nash	H_2 (t)	212.03 ± 53.29	217.45 ± 65.31	−0.28	0.782	1.000	−0.09	[−40.7, 33.7]	0.184
	Revenue (M€)	1.216 ± 0.040	1.217 ± 0.042	−0.06	0.956	1.000	−0.02	[−0.026, 0.025]	0.919
H-TD3 vs. TD3	H_2 (t)	212.03 ± 53.29	214.19 ± 55.67	−0.12	0.904	1.000	−0.04	[−34.7, 33.4]	0.313
	Revenue (M€)	1.216 ± 0.040	1.238 ± 0.049	−1.55	0.130	0.907	−0.50	[−0.049, 0.007]	0.077

Following the decision rule that conclusion strength is governed jointly by the Holm-corrected p and the bootstrap CI (with the CI taking precedence under unequal n), the contrasts read as follows. TD3 was directionally higher than RBC in revenue (+0.033 M€;

the bootstrap CI excludes zero, and the MWU test rejects equality, $p < 0.001$, but the Holm-corrected Welch test does not, $p = 0.105$)—the largest revenue effect in the ablation, paired with the 1/18 collapse exposure of the same configuration. H-TD3 and RBC were statistically indistinguishable in revenue (+0.010 M€; CI [−0.007, 0.028] M€; $g = 0.34$). RBC produced directionally more hydrogen than H-TD3 (34.6 t; CI excludes zero; MWU $p = 0.068$ and Holm $p = 0.090$, both marginal; the same pattern holds for TD3, −32.5 t) and a significantly higher biomass gain (5.7 percentage points; Holm $p = 0.001$; CI [−7.78, −3.68]; $g = -1.62$)—the only Holm-confirmed mean effect in the family, and it favors the baseline. All log-Nash-internal contrasts (TD3+Nash vs. TD3; H-TD3 vs. TD3+Nash; H-TD3 vs. TD3) were statistically indistinguishable in hydrogen and revenue (all CIs include zero; $|g| \leq 0.50$).

For power analysis, the contrasts against RBC with $|g| \geq 0.83$ required 6–23 seeds per group at 80% power and were accordingly detectable at the available sample sizes; the TD3-vs-RBC revenue contrast ($g = 0.94$, required $n \approx 18$) sits exactly at the detection boundary, which is why its CI excludes zero while the Holm-corrected Welch test does not. The revenue contrasts among RL configurations ($|g| \leq 0.50$) would require 63 seeds per group or more (up to tens of thousands for the near-null contrasts), and the H-TD3-vs-RBC revenue contrast ($g = 0.34$) $\approx$ 132 seeds per group. Mean-effect claims below the detectable range are accordingly avoided throughout; the collapse analysis of Section 5.4 remains the appropriate lens for the reward-structure comparison because collapse is a threshold event detectable at these sample sizes.

5.6. Training Convergence

Convergence is measured on the equal-weight physical objective $\bar{J} = \left(J_{H2} + J_{fish} + J_{eco}\right)/3$, which has an identical definition across reward designs and is therefore comparable between the scalarized and log-Nash families, unlike raw training rewards (Section 5.3). Two descriptive statistics are reported per configuration: (i) the time to 95%—the first episode at which the 50-episode rolling mean of $\bar{J}$ reaches 95% of its final value—and (ii) the $\bar{J}$-plateau—the first episode from which this rolling mean remains within ±0.5% of its final value. H-TD3 reached 95% of its final $\bar{J}$ at episode 50.0 ± 0.0 in every one of its 20 seeds (one episode = 96 h, i.e., ≈4800 simulated training hours)—the earliest observable episode given the window length, with zero seed dispersion—while TD3, TD3+PIRL, TD3+Hier, and TD3+Nash averaged 62 ± 37, 65 ± 44, 53 ± 8.5, and 78 ± 55 episodes (medians 50). The $\bar{J}$-plateaus were reached at 220 ± 71 (TD3), 201 ± 63 (TD3+PIRL), 264 ± 72 (TD3+Hier), 270 ± 77 (TD3+Nash), and 212 ± 8.4 (H-TD3) episodes; the H-TD3 plateau location is statistically indistinguishable from TD3 (Mann–Whitney p = 0.85), while its cross-seed dispersion is an order of magnitude smaller (Levene test: $p = 1 \times 10^{-4}$ vs. TD3; $p = 2 \times 10^{-4}$ vs. TD3+Nash). Because the early-stopping horizon bounds the number of training episodes (220 for all H-TD3 runs, up to 400 for slower configurations), plateau values near the horizon are weakly right-censored; the time-to-95% statistic is unaffected by this censoring. This subsection also corrects the convergence statistics of the previous version, which had been computed on raw training rewards with a one-sided first-crossing criterion; that criterion did not require the rolling mean to remain at its final level and was blind to the definitional downward drift of the log-Nash reward during the disagreement-point annealing of the first 120 episodes (Equation (19)), which understated the H-TD3 plateau episodes. Figure 4 shows the per-seed mean ± SD bands of the 50-episode rolling mean of $\bar{J}$; the ordinate is common to both panels.

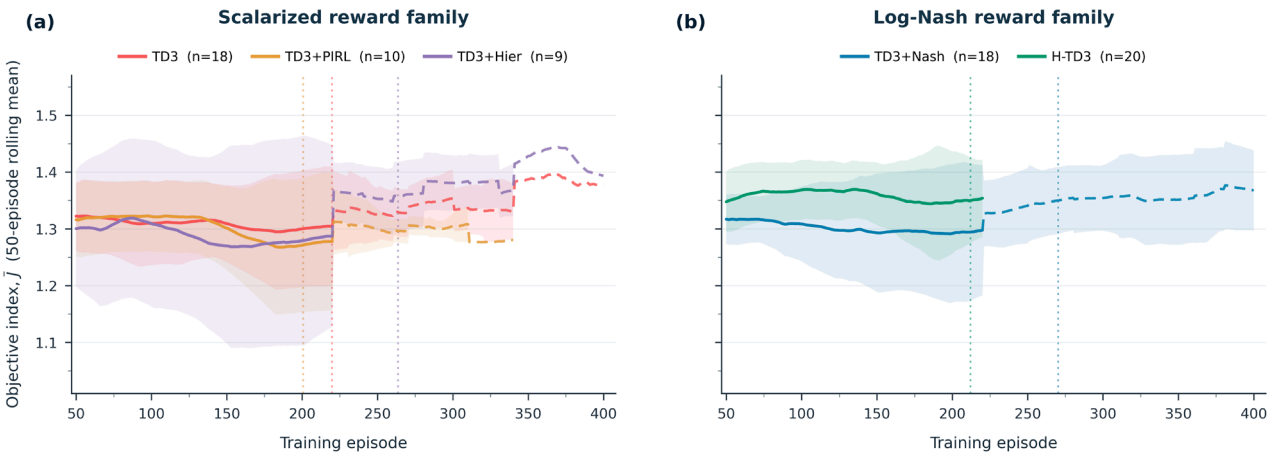


Figure 4. Training convergence in the physical-objective space: 50-episode rolling mean of the equal-weight objective $\bar{J} = \left(J_{H2} + J_{fish} + J_{eco}\right)/3$ (line = mean over seeds, band = ±1 SD; dotted vertical lines = per-configuration mean $\bar{J}$-plateau episode). (**a**) Scalarized reward family (TD3, TD3+PIRL, TD3+Hier; Equation (20)); (**b**) log-Nash family (TD3+Nash, H-TD3; Equation (17)). $\bar{J}$ has an identical definition across reward designs, so the ordinate is common to both panels; the rolling window matches the plateau criterion of Section 5.6.

5.7. Time-Domain Behavior over the 1000 h Horizon

Figure 5 presents the recorded 1000 h evaluation time series for H-TD3 (seed 0) and the rule-based controller (RBC) under identical environmental and electricity-price realizations. Compared with the power-following behavior of RBC, H-TD3 generally operates the electrolyzer at a lower power level, particularly during low-price periods. The average electrolyzer power was 12.56 MW for H-TD3, compared with 14.15 MW for RBC. Although the normalized electricity price repeatedly reached its upper cap of 1.0 (equivalent to €120/MWh), the recorded dispatch did not exhibit a systematic increase in electricity exports during price spikes. Nevertheless, H-TD3 achieved a higher cumulative electricity export of 5101 MWh, compared with 3492 MWh for RBC, while producing 222.4 t of hydrogen versus 246.1 t for RBC.

Regarding operational constraints, the cage temperature under H-TD3 ranged from 12.0 to 17.94 °C, exceeding the upper boundary of the 12–16 °C comfort band for approximately 965.6 h, whereas the maximum temperature under RBC was 14.51 °C. Dissolved oxygen remained above the 6.0 mg/L hard safety limit under both policies, with minimum concentrations of 6.532 and 6.488 mg/L for H-TD3 and RBC, respectively. While H-TD3 maintained dissolved oxygen above the 6.5 mg/L barrier-activation threshold, RBC experienced approximately 20.1 h below this threshold. The final electrolyzer stack state of health remained approximately 0.9872 for both policies, and the stack temperature remained below the modeled 95 °C safety ceiling.

The baseline environmental realization exhibited a maximum wind speed of 15.93 m/s and a significant wave height of 3.45 m. Consequently, this evaluation does not exercise the turbine cut-out regime (25 m/s) or severe-sea conditions, which are examined separately in the dedicated 72 h storm scenario presented in Figure 6.

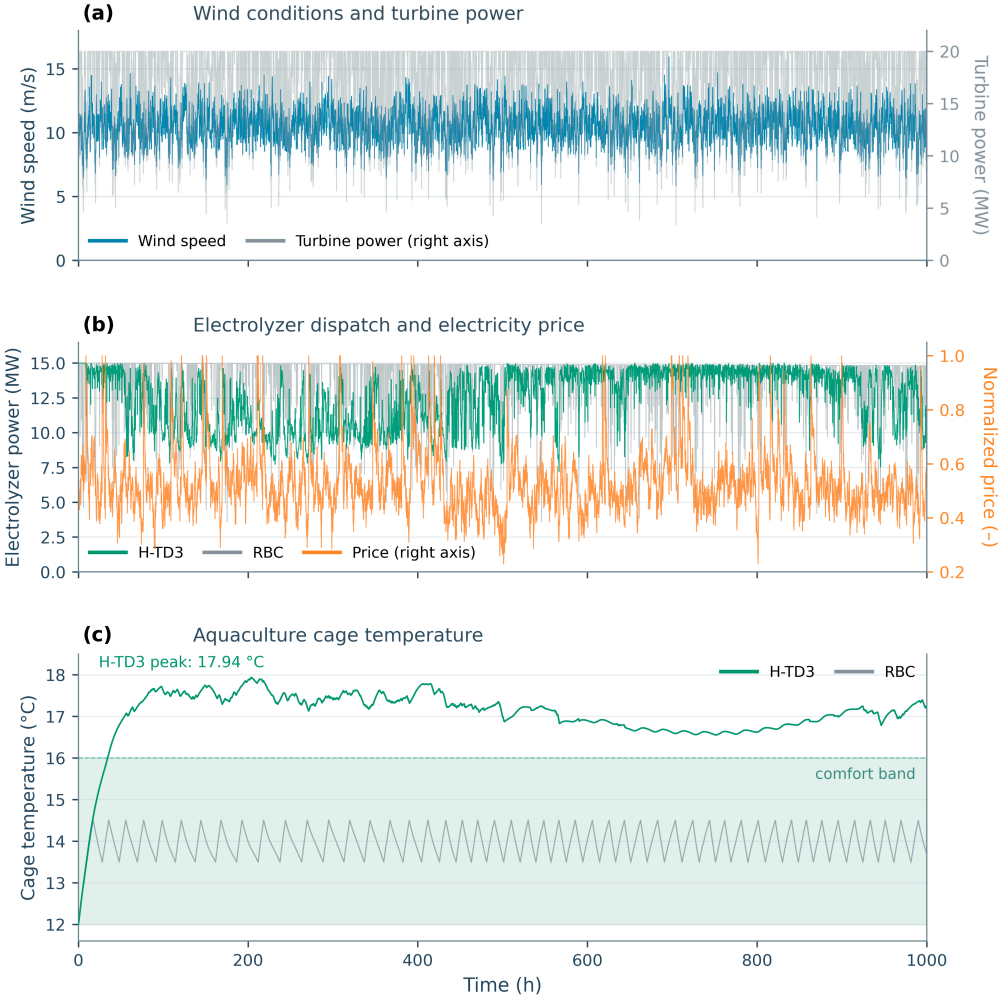


Figure 5. Representative 1000 h evaluation time series comparing H-TD3 (seed 0, green) and RBC (seed 0, grey): (**a**) wind speed and turbine power (right axis); (**b**) electrolyzer power and normalized electricity price (right axis, orange); and (**c**) cage temperature relative to the 12–16 °C comfort band. Both policies are evaluated under identical wind and electricity-price realizations. The H-TD3 trajectory exceeds the upper comfort threshold for much of the evaluation period, highlighting a trade-off between energy dispatch and aquaculture temperature regulation.

5.8. Extreme Sea-State Scenario (72 h)

The baseline 1000 h windows of Section 5.7 contain no cut-out events (wind $\leq$ 15.9 m/s, $H_s \leq$ 3.45 m). A dedicated 72 h Beaufort-10 storm scenario was therefore run to exercise the full environmental envelope of Equation (1), including the cut-out logic and the load–pitch coupling of Equation (4) at high sea states: wind ramps from 12 to 26.5 m/s over 6 h, holds a 54 h storm plateau around the 25 m/s cut-out with intermittent turbine shutdown/restart (50.2 of 72 h above cut-out; turbine availability 30.3%), and decays to 14 m/s over 12 h; significant wave height reaches a platform-phase mean of 7.6 m with a 9.1 m peak. All six configurations were evaluated on the same meteorological realization (scenario seed 4242), so the differences below are policy behaviors under identical forcing. This is a mechanism demonstration under a single scenario, not a statistical comparison; no significance claims are made or implied. Table 6 summarizes the per-configuration outcomes.

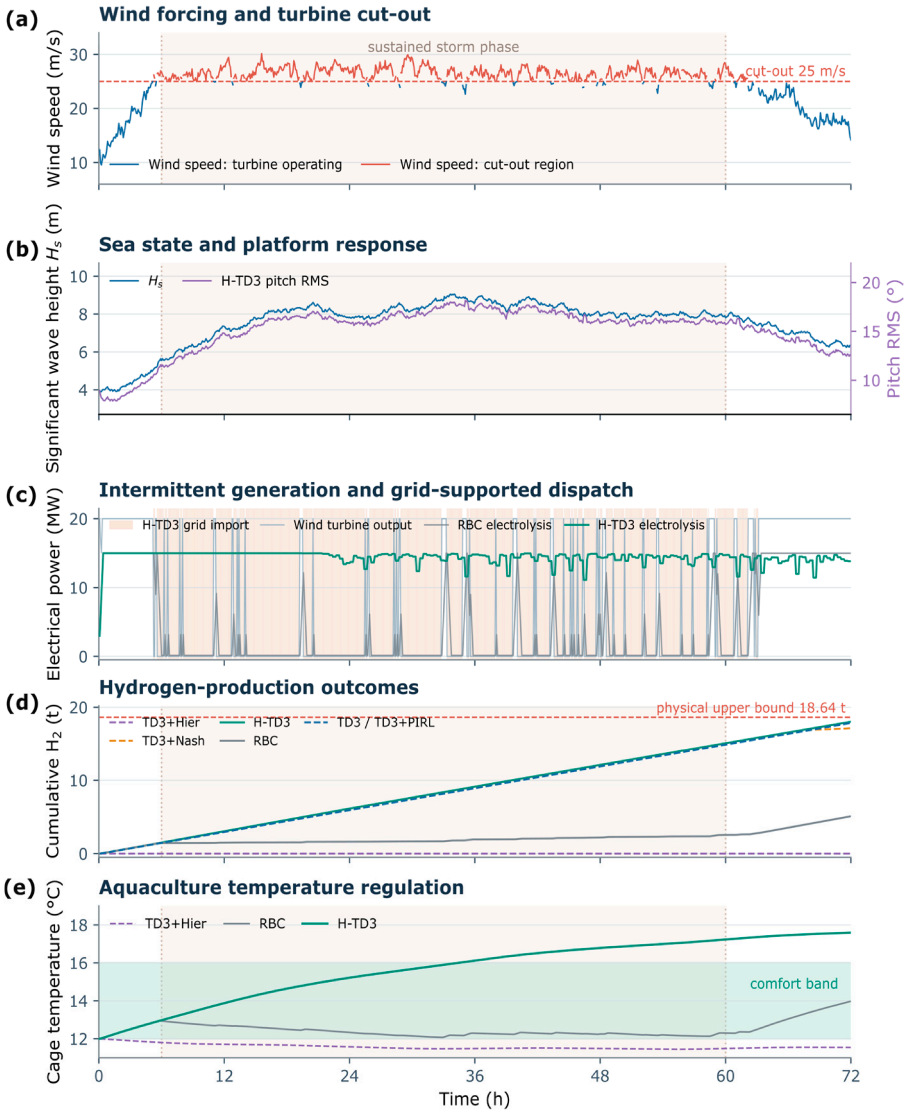


Figure 6. 72 h Beaufort-10 storm scenario (single meteorological realization, seed 4242), with all configurations evaluated under identical forcing. (**a**) Wind speed and turbine operating status: the blue line shows wind speed, the red segments indicate periods when wind speed exceeds the turbine cut-out threshold, and the horizontal dashed line marks the 25 m/s cut-out wind speed. (**b**) Significant wave height (left axis) and platform pitch RMS (right axis). (**c**) Electrolyzer power. (**d**) Cumulative hydrogen production compared with the 18.64 t physical ceiling for 72 h. (**e**) Cage temperature with the comfort band indicated by the shaded region. In panels (**a**,**b**,**d**,**e**), the light shaded background denotes the sustained storm phase, and the vertical dotted lines indicate the transitions into and out of this phase. TD3+PIRL is bit-identical to TD3 and is omitted for legibility. The collapsed TD3+Hier policy (seed 0) holds the electrolyzer at standby throughout; RBC idles through the cut-out gaps; the learned policies buffer hydrogen across intermittent availability windows.

Three observations follow. First, the learned policies maintain hydrogen production through grid-supported operation during wind unavailability, whereas power-following idles: H-TD3 produced 18.03 t, i.e., 96.7% of the 72 h physical ceiling (18.64 t at the model's 68% HHV rating), versus 5.09 t (27.3%) for RBC—a 3.5× difference. This is not a more complete conversion of the scarce wind: of the 1037 MWh consumed by the electrolyzer under H-TD3, 729 MWh (about 70%) was imported from the grid, versus 58 MWh of 298 MWh (about 19%) for RBC, so wind supplied only about 30% (308 MWh) of the H-TD3 electrolyzer energy. Because imported electricity is not assumed to be renewable (Section 2.5), most of this storm hydrogen is grid-supported rather than green hydrogen in

the strict sense. The mechanism is hydrogen-buffered, grid-supported dispatch: the learned policy holds the electrolyzer near its feasible maximum across the intermittent available-power windows, drawing grid import to cover the calm and cut-out windows instead of letting these gaps idle the plant; the zero-production segments during cut-out are clearly visible in the time history (Figure 6). Second, the storm is precisely the regime in which the collapse pathology of Section 5.4 manifests most starkly: the collapsed TD3+Hier seed (seed 0) produced zero hydrogen over the full 72 h and exported the scarce storm wind, while the non-collapsed log-Nash and H-TD3 policies produced 17.1–18.0 t under identical forcing. Third, the trade-offs are reported transparently: under heat-priority reallocation the high-electrolyzer-load policies pushed the mean cage temperature to 17.6 °C, above the 12–16 °C comfort band, and consequently RBC—the lowest electrolyzer load policy, whose cage stays close to its natural temperature—achieved the highest storm biomass gain (+1.61% vs. +1.42% for H-TD3). Pitch RMS maxima (H-TD3 18.1° vs. RBC 20.9°) are reported as differences of the linearized load–response proxy of Equation (4) under different dispatch; these values lie far outside the OpenFAST-verified range (pitch RMS $\leq 3.8°$; $H_s \leq 5.7$ m in the P50 family at 25 m/s, where the proxy gives 11.6–13.5° against an OpenFAST pitch standard deviation of 0.65°, versus a 9.1 m peak H_s in the storm) and are linear extrapolations of a conservative index, not physical pitch predictions; the decomposition and sensitivity analysis of Figure S7 quantify how much of such differences the load-coupling term can account for. Conservation residuals remained at machine precision throughout the storm (max 3.1×10^{-16} per step), minimum dissolved oxygen never fell below 6.59 mg/L, and all safety constraints were respected in every configuration. Storm-window revenues are small in absolute terms (€25.7–36.3 k over 72 h across configurations).

Table 6. Seventy-two hour storm scenario, single meteorological realization (seed 4242). Mechanism demonstration only.

Configuration	H_2 Produced (t)	Share of 18.64 t Ceiling	Electrolyzer Energy (MWh)	Pitch-Proxy RMS Max (°; Extrapolated)	Cage T Mean (°C)	Biomass Gain (%)
RBC	5.09	27.3%	298	20.9	14.0	1.61
TD3 = TD3+PIRL	17.85	95.8%	1025	18.1	17.6	1.42
TD3+Hier	0.00	0.0%	11	20.9	11.5	1.36
TD3+Nash	17.14	92.0%	987	19.2	16.9	1.44
H-TD3	18.03	96.7%	1037	18.1	17.6	1.42

For load–response attribution, because the pitch proxy of Equation (4) embeds the electrolyzer load through the relief term $-\mathrm{KL}\cdot\ell$, pitch differences between dispatch policies mix wave forcing with a policy-dependent load effect. A decomposition of the 1000 h evaluation (Figure S7) shows that the wave-driven term dominates for both RBC and H-TD3, and a KL sensitivity sweep (0.05–0.25) bounds the entire load effect between the two policies at $\leq 0.76°$ ($\approx$15% of the mean). The pitch differences in Table 3 and in the storm scenario are therefore small perturbations on a wave-dominated response, and no operational-safety claim is attached to them.

5.9. PIRL Ablation: A Negative Result

The physics-informed residual term had no measurable effect, and the root cause has been established: in the conservation-closed simulator, the penalty is structurally zero at machine precision. Under a random policy, the four conservation residuals never exceeded 3.4×10^{-16} per step, about eight orders of magnitude below the float32 representation step at the typical reward magnitude; in a same-seed short-training comparison, the float32 rewards of TD3 and TD3+PIRL agreed bit-for-bit and the trained actor and critic parameters

were identical (matching SHA-256 hashes); and flipping the sign of the residual term changed no reward (full diagnostic in Supplementary Section S9). Consistently, the five original TD3+PIRL seeds reproduced plain TD3 bit-for-bit, and five additional independent seeds (10 in total; Table 3) showed the same distributional behavior as the scalarized family, with a single collapse (seed 4, the counterpart of the collapsed TD3 seed). Because bit-identical replicates are not independent observations, TD3+PIRL is treated as a diagnostic confirmation of TD3 and is excluded from the pooled collapse statistics of Section 5.4 (Table 4, Figure 3). No PIRL ablation effect is claimed: residuals closed to machine precision carry no gradient information, so the term cannot influence training at any weight, and PIRL is repositioned as the numerical integrity audit of Section 3.5. Whether a physics-informed term can add value in a simulator whose residuals are *not* closed by construction is a question this test cannot answer.

6. Techno-Economic Assessment

6.1. Assumptions and LCOH

All monetary quantities are in euros. The assessment covers the electrolysis plant (the capital expenditure (CAPEX) of the wind farm, and the aquaculture system lies outside the assessment boundary). The owned-wind result is therefore an electrolyzer-subsystem (ex-wind-electricity) cost rather than a full-system LCOH. The assumptions are listed in Table S8 (plant CAPEX €1200/kW, stack replacement at 35% of CAPEX, operation and maintenance (O&M) 3% of CAPEX per year, 6% discount rate over 25 years, 8000 equivalent operating hours per year, hydrogen price €4.0/kg, waste-heat credit €0.05/kWh_th); the levelized cost of hydrogen is computed on two bases:

$$\begin{cases} LCOH_{ex\text{-}elec} = \frac{CAPEX \cdot CRF(0.06,25) + A_{stack} + O\&M}{M^{yr}_{H_2}} \\ LCOH_{incl\text{-}elec} = LCOH_{ex\text{-}elec} + \frac{C^{yr}_{elec}}{M^{yr}_{H_2}} \end{cases} \quad (23)$$

where CRF is the capital recovery factor, A_{stack} the annualized stack-replacement cost (replacement cost 35% of plant CAPEX, amortized linearly over the stack life of 15,643 equivalent full-load hours extrapolated from the simulated SoH slope of 1.279%/1000 h), and production is annualized from the 1000 h simulation at 8000 equivalent hours/year. The resulting LCOH and project economics are summarized in Table 7.

Table 7. LCOH and project economics. "Owned wind" basis: electricity supplied internally by the wind farm (its CAPEX accounted separately at the wind-farm level, the configuration relevant for a multi-use FOWT platform; no electricity cost or export opportunity cost is charged, so the result is an electrolyzer-subsystem, ex-electricity cost that excludes wind-farm and aquaculture CAPEX); "grid purchase" basis: electricity bought at €60/MWh.

Quantity	Value
Electrolyzer-subsystem cost (ex-electricity), owned-wind basis (H-TD3 production)	€3.05/kg
—of which stack replacement amortization	€1.90/kg
Electrolyzer-subsystem cost (ex-electricity), owned-wind basis (RBC production, reference normalization)	€2.62/kg
LCOH including electricity valued at €60/MWh (grid purchase or export-price opportunity cost)	€6.45/kg (electricity share €3.40/kg)
NPV, owned-wind basis (electrolyzer subsystem only; zero-cost wind electricity)	+€23.4 M; IRR 18.4%
NPV, grid-purchase basis (electricity valued at €60/MWh)	−€50.2 M; IRR undefined (negative result, reported as such)

Three observations follow. First, the owned-wind figure of €3.05/kg is an electrolyzer-subsystem (ex-electricity) cost: it excludes wind-farm and aquaculture CAPEX and assigns

no cost to the wind electricity, so it should not be compared directly with full-boundary LCOH values from the literature (e.g., €4.01/kg for a 250 MW degradation-aware fleet [26]) unless the system boundaries are equivalent. Second, the owned wind has an opportunity cost because it could otherwise be exported: valuing it at the simulation-average price of €60/MWh (a flat-price approximation; dispatch-weighted prices would be lower because the learned policies shift electrolysis to low-price periods) reproduces the grid-purchase basis, which is *not* viable at €60/MWh: purchasing electricity at the simulation-average price destroys value (NPV −€50.2 M). We report the negative case deliberately because it delimits the electrolyzer-subsystem economics: they close only if the electricity carries a negligible opportunity cost, and no conclusion on the economic viability of the integrated wind–hydrogen–aquaculture platform is drawn, because wind-farm and aquaculture CAPEX and revenues lie outside the boundary. Third, the €2.62/kg figure is a pure production-normalization of the same fixed-cost base (fixed costs per kilogram scale inversely with annual production; it uses the RBC output of 1973.4 t/yr with identical CAPEX/O&M/stack assumptions) and is provided for cross-study comparability, not as a projected improvement.

6.2. Sensitivity

The dominant lever is stack lifetime: tripling stack life removes €1.27/kg (ex-electricity), which is why the durability question—unresolvable within a 1000 h reduced-order proxy (Section 7)—matters more for this concept than any dispatch refinement quantified here. Electricity price is the second lever on the purchased-power basis and irrelevant on the owned-wind basis only because that boundary assigns the wind electricity no cost. CAPEX scaling shifts all fixed components proportionally, O&M is a minor lever at ±3% of CAPEX, and the production band (±1 SD across the 20 H-TD3 seeds, ±53.3 t/1000 h) spans €2.44–4.07/kg ex-electricity—policy-seed variability alone is worth as much as a ±20% CAPEX swing, which quantifies why collapse avoidance matters economically and not only operationally. Figure 7 visualizes both bases with all five levers (numerical values in Table S9). Oxygen co-production and illustrative carbon-avoidance figures are given in Supplementary Section S10.

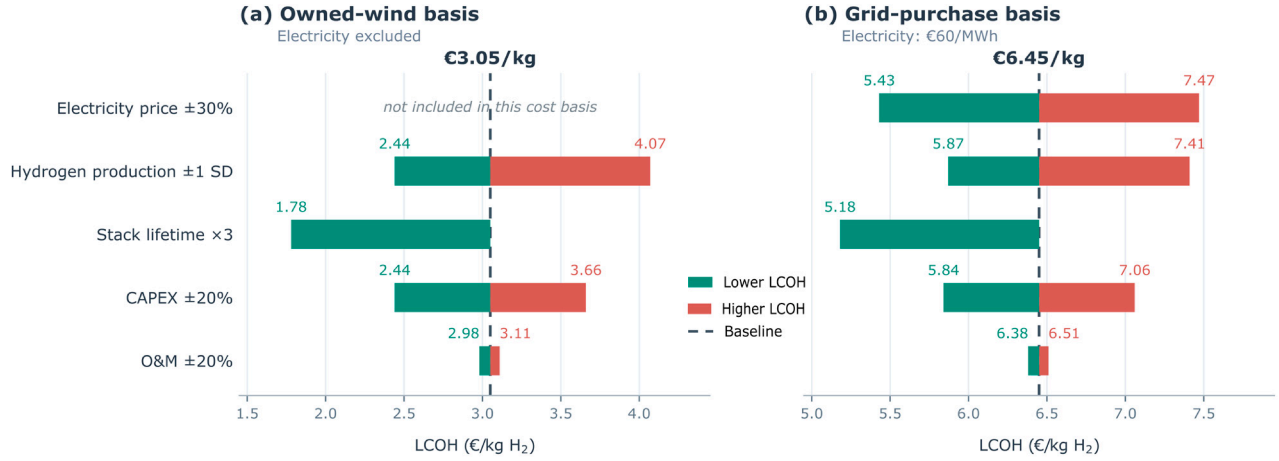


Figure 7. LCOH sensitivity (tornado diagram). (**a**) Owned-wind, ex-electricity basis (electrolyzer-subsystem cost; base case €3.05/kg); (**b**) grid-purchase basis at €60/MWh (base case €6.45/kg). Bars span the LCOH range of each lever: CAPEX ±20%, O&M ±20%, stack life ×3 (15,643 → 46,928 equivalent full-load hours), electricity price ±30% (€42–78/MWh), and production ±1 SD (policy-seed variability of the 20 H-TD3 runs). The vertical dashed lines indicate the baseline LCOH values; green and red bars represent lower and higher LCOH values, respectively. Stack lifetime dominates both bases; the production band is the only policy-controllable lever and is co-equal with CAPEX on the owned-wind basis.

7. Conclusions and Limitations

This study developed a conservation-closed reduced-order simulation framework to investigate hierarchical multi-objective dispatch of a 20 MW floating wind–hydrogen–aquaculture platform with OpenFAST-informed environmental modeling. The proposed framework ensures numerical consistency through closed-loop mass, energy, and thermodynamic balance verification, providing an auditable basis for evaluating integrated offshore energy systems.

The main contribution of the proposed annealed log-Nash formulation (H-TD3) is improved training robustness rather than increased mean revenue or reduced performance variance. Across independent training runs, H-TD3 avoided the zero-hydrogen shutdown collapse observed in scalarized and fixed-disagreement Nash baselines and achieved markedly more consistent convergence ($\bar{J}$-plateau dispersion an order of magnitude below the scalarized baselines; Section 5.6). Within the tested seeds, these results suggest that the log-Nash formulation may mitigate reward-conflict-induced failure modes, while the current sample size does not establish a statistically significant difference in collapse probability. The benefit of H-TD3 should therefore be interpreted as an empirical robustness improvement rather than a guaranteed statistical advantage.

Compared with a strong rule-based controller, H-TD3 achieved statistically comparable economic performance while satisfying operational constraints. The learned policies mainly improved adaptive energy allocation, shifting the balance among electricity export, hydrogen production, and aquaculture services rather than simply maximizing revenue. In contrast, plain TD3 achieved higher average revenue but exhibited occasional collapse, revealing a trade-off between aggressive optimization and operational robustness. The economic analysis further shows that the electrolyzer subsystem breaks even only when the electricity carries a negligible opportunity cost (owned wind, excluding wind-farm and aquaculture CAPEX); at the €60/MWh average price, it destroys value (NPV −€50.2 M), and the viability of the integrated platform is not assessed. Stack lifetime and production variability are major determinants of the electrolyzer-side hydrogen cost. Therefore, robust control contributes economic value by reducing operational failure exposure, not only by improving short-term performance.

Several limitations should be acknowledged. First, the degradation model is a reduced-order representation calibrated through empirical coefficients; because state-of-health variation is negligible within the short RL training horizon, the controller does not learn life-cycle durability trade-offs, which are considered only through replacement-cost modelling. Second, platform motion is represented by a linearized response model with $\zeta_M = 1$ in the dispatch layer. External OpenFAST verification and experimental evidence provide bounded support for this assumption, but full-scale 20 MW aero-hydro-electrochemical coupling and mechanistic motion-induced degradation remain unresolved. Platform pitch is carried only as a one-way diagnostic proxy: it is not part of the state, reward, electrolyzer, or degradation models, it overpredicts the OpenFAST pitch-RMS envelope (5.0–5.2° vs. 0.6–3.8°), and its storm values lie outside the verified range; the present controller is therefore not motion-response-aware in the sense of exploiting platform motion in electrolyzer operation. Third, the physics-informed residual term currently serves primarily as a numerical integrity check because the conservation residuals of the simulator are already near machine precision; no additional control improvement from PIRL is claimed. Fourth, environmental forcing and aquaculture dynamics remain simplified, relying on synthetic stochastic inputs and literature-based biological parameters rather than measured site-specific data. Fifth, grid imports (2–4% of electrolyzer energy in the 1000 h evaluations and about 70% in the storm scenario) are not assumed to be renewable, so hydrogen produced from imports is

grid-supported rather than green, and the techno-economic assessment is restricted to the electrolyzer subsystem.

Future work will focus on improving physical and operational fidelity through three directions: (i) ensemble-based extreme-weather evaluation and measured-site environmental forcing; (ii) higher-fidelity aero-hydro-electrochemical coupling, including 20 MW-specific OpenFAST models, explicit platform-motion-dependent electrolyzer performance and degradation in the state and dispatch decision, and improved degradation mechanisms; and (iii) larger-scale statistical validation with expanded training seeds to further assess the robustness of hierarchical control strategies.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/en19225521/s1, Figure S1: PEM electrolyzer model characteristics; Figure S2: Aquaculture cage model; Figure S3: Reward and safety design; Figure S4: Anchoring of the environmental processes to the ERA5 reanalysis (2014–2023, three North Sea sites, hourly); Figure S5: Conservation audit (seed 777, RBC policy, 1000 h); Figure S6: Ablation distributions over seeds (1000 h deterministic evaluation); Figure S7: Decomposition of the linearized pitch proxy (Equation (4), 1000 h evaluation, seed 0). Table S1: Implemented system parameters (all values as used in code); Table S2: PEM electrochemical and degradation parameters with provenance; Table S3: Heat-generation reference sensitivity (counterfactual replay, 1000 h deterministic evaluation, mean ± SD over seeds); Table S4: Hyperparameters (as implemented); Table S5: Stochastic environment processes; Table S6: Parameter anchoring to the ERA5 reanalysis; Table S7: OpenFAST verification campaign and proxy comparison; Table S8: Techno-economic assumptions (€-denominated); Table S9: LCOH sensitivity.

Author Contributions: Conceptualization, M.L. and H.H.; methodology, M.L. and C.F.; software, H.H.; validation, L.Y., X.W., C.F. and X.Z.; formal analysis, M.L.; investigation, H.H., L.Y. and X.Z.; resources, H.H.; data curation, X.W.; writing—original draft preparation, M.L.; writing—review and editing, H.H. and C.F.; visualization, M.L.; supervision, M.L.; project administration, M.L.; funding acquisition, M.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Science and Technology Major Project of China, "100 MW-level Offshore Integrated Energy Island Demonstration Project", grant number 2024ZD0801704.

Data Availability Statement: The complete computational package supporting the conclusions of this study-simulation and reinforcement-learning source code, the seed protocol (Table S4), all trained actor weights, per-seed evaluation results and statistical outputs, and the 90-case OpenFAST verification data (case matrix, batch statistics, and per-case proxy-error tables) underlying Table S7 and the collapse analysis of Section 5.4—is provided as Supplementary Materials (ZIP archive), permanently accessible via the article's DOI page. No proprietary components are involved; all external tools (OpenFAST v4.2.1, ROSCO v2.9.0, the IEA 15 MW reference turbine, and the UMaine VolturnUS-S platform) are open-source and publicly licensed.

Acknowledgments: The authors are profoundly grateful to the Editor for handling this manuscript and to the anonymous reviewers for their invaluable comments and constructive suggestions, which significantly improved the quality of this work. During the preparation of this work, the authors used ChatGPT (GPT-6; OpenAI, San Francisco, CA, USA) to improve the clarity of the writing and assist in the preparation of Figure 1. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the final content of the publication.

Conflicts of Interest: Authors M.L., C.F., H.H., L.Y., and X.W. were employed by the company PowerChina Huadong Engineering Corporation Limited. Authors M.L., C.F., and H.H. were also employed by the company PowerChina Jiangsu Engineering Corporation Limited. Author X.Z. was employed by the company CTG Jiangsu Energy Investment Co., Ltd. The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Díaz, H.; Guedes Soares, C. Review of the Current Status, Technology and Future Trends of Offshore Wind Farms. *Ocean Eng.* **2020**, *209*, 107381. [CrossRef]
2. Barooni, M.; Ashuri, T.; Velioglu Sogut, D.; Wood, S.; Ghaderpour Taleghani, S. Floating Offshore Wind Turbines: Current Status and Future Prospects. *Energies* **2023**, *16*, 2. [CrossRef]
3. Singlitico, A.; Østergaard, J.; Chatzivasileiadis, S. Onshore, Offshore or In-Turbine Electrolysis? Techno-Economic Overview of Alternative Integration Designs for Green Hydrogen Production into Offshore Wind Power Hubs. *Renew. Sustain. Energy Transit.* **2021**, *1*, 100005. [CrossRef]
4. Carmo, M.; Fritz, D.L.; Mergel, J.; Stolten, D. A Comprehensive Review on PEM Water Electrolysis. *Int. J. Hydrogen Energy* **2013**, *38*, 4901–4934. [CrossRef]
5. Buck, B.H.; Langan, R. (Eds.) *Aquaculture Perspective of Multi-Use Sites in the Open Ocean*; Springer: Cham, Switzerland, 2017.
6. Miranda, F.; Rosa-Santos, P.; Taveira-Pinto, F.; Guan, D.; Fazeres-Ferradosa, T. Aquaculture and Offshore Wind: A Review of Co-Location Design Challenges. *Ocean Eng.* **2025**, *318*, 120161. [CrossRef]
7. Maar, M.; Holbach, A.; Boderskov, T.; Thomsen, M.; Buck, B.H.; Kotta, J.; Bruhn, A. Multi-Use of Offshore Wind Farms with Low-Trophic Aquaculture Can Help Achieve Global Sustainability Goals. *Commun. Earth Environ.* **2023**, *4*, 447. [CrossRef]
8. Zhao, Z.; Wu, W.; Lin, R.; Liu, Y. Polarization Characteristics of an Alkaline Water Electrolyzer under Marine Sloshing Conditions. *J. Mar. Sci. Eng.* **2026**, *14*, 660. [CrossRef]
9. Choi, Y.; Lee, W.; Na, Y. Effect of Gravity and Various Operating Conditions on Proton Exchange Membrane Water Electrolysis Cell Performance. *Membranes* **2021**, *11*, 822. [CrossRef] [PubMed]
10. Chen, J.; Sun, Y.; Pan, X.; Tao, A.; Jia, X.; Zhang, C.; Lv, H. Effects of Gravity on the Cell Performance of Proton Exchange Membrane Water Electrolyzer. *J. Alloys Compd.* **2025**, *1010*, 177884. [CrossRef]
11. Zhu, D.; Xu, L.; Su, X.; Hu, B.; Jia, T.; Mi, L. Experimental Study of the Effect of Mechanical Vibration and Water Velocity on Bubble Management in PEM Electrolysis Cell. *Int. J. Hydrogen Energy* **2024**, *49*, 390–403. [CrossRef]
12. Purnami, P.; Satrio, N.W.; Sofi'i, Y.K.; Wardana, I.N.G. The Impact of Mechanical Vibration at Cathode on Hydrogen Yields in Water Electrolysis. *J. Power Sources* **2024**, *615*, 235075. [CrossRef]
13. Li, Z.-H.; Wu, D.; Hu, K.-X. Detachment of Electrolytic Bubbles in Inclined Channels. *Int. J. Hydrogen Energy* **2026**, *230*, 154813. [CrossRef]
14. Matsushima, H.; Nishida, T.; Konishi, Y.; Fukunaka, Y.; Ito, Y.; Kuribayashi, K. Water Electrolysis under Microgravity: Part 1. Experimental Technique. *Electrochim. Acta* **2003**, *48*, 4119–4125. [CrossRef]
15. Liu, T.; Zhao, Z.; Tang, W.; Chen, Y.; Lan, C.; Zhu, L.; Jiang, W.; Wu, Y.; Wang, Y.; Yang, Z.; et al. In-Situ Direct Seawater Electrolysis Using Floating Platform in Ocean with Uncontrollable Wave Motion. *Nat. Commun.* **2024**, *15*, 5305. [CrossRef] [PubMed]
16. Yang, L.; Zhang, K.; Chen, X.; Chen, Y.; Li, P. Flow Separation Characteristics of Hydrogen and Electrolyte in T-Junction Separators of Floating Wind Power Hydrogen Production Systems. *Int. J. Hydrogen Energy* **2025**, *168*, 151050. [CrossRef]
17. Pham, T.D.; Dinh, V.N.; Trinh, L.C.; Judge, F.; Leahy, P. Dynamic Assessment of Decentralized Hydrogen Production and Storage on a Semi-Submersible Floating Wind Turbine Platform. *Int. J. Hydrogen Energy* **2025**, *164*, 150832. [CrossRef]
18. Wang, H.; He, S.; Yan, J.; Han, S.; Liu, Y. Deep Reinforcement Learning-Driven Wind Farm Flow Control Considering Dynamic Wind. *Energy Convers. Manag.* **2025**, *337*, 119888. [CrossRef]
19. Jendoubi, I.; Bouffard, F. Multi-Agent Hierarchical Reinforcement Learning for Energy Management. *Appl. Energy* **2023**, *332*, 120500. [CrossRef]
20. Peters, H.J.M. Nash Bargaining Solutions. In *Axiomatic Bargaining Game Theory*; Springer: Dordrecht, The Netherlands, 1992; pp. 13–45. [CrossRef]
21. Wang, Y.; Zheng, Y.; Yang, Q. Nash Bargaining Based Collaborative Energy Management for Regional Integrated Energy Systems in Uncertain Electricity Markets. *Energy* **2023**, *269*, 126725. [CrossRef]
22. Li, N.; Wang, G.; Guo, D.; Pan, C. Optimal Energy Sharing Strategy in Multi-Integrated Energy Systems Considering Asymmetric Nash Bargaining. *Energies* **2025**, *18*, 5729. [CrossRef]
23. Hersbach, H.; Bell, B.; Berrisford, P.; Hirahara, S.; Horányi, A.; Muñoz-Sabater, J.; Nicolas, J.; Peubey, C.; Radu, R.; Schepers, D.; et al. The ERA5 Global Reanalysis. *Q. J. R. Meteorol. Soc.* **2020**, *146*, 1999–2049. [CrossRef]
24. *IEC 61400-1:2005*; Wind Turbines—Part 1: Design Requirements. International Electrotechnical Commission: Geneva, Switzerland, 2005.
25. Maoulida, F.; Guilbert, D.; Camara, M.-B.; Dakyo, B. Dynamic Electrical Degradation of PEM Electrolyzers under Renewable Energy Intermittency: Mechanisms, Diagnostics, and Mitigation Strategies—A Comprehensive Review. *Renew. Sustain. Energy Rev.* **2026**, *225*, 116170. [CrossRef]
26. Nezhadkhatami, H.; Hajizadeh, A.; Soltani, M. Techno-Economic Optimization of Power Allocation in Large-Scale Hydrogen Production with Degradation-Aware Control Strategies. *Int. J. Hydrogen Energy* **2026**, *203*, 153113. [CrossRef]

27. Zhang, H.; Yuan, T. Optimization and Economic Evaluation of a PEM Electrolysis System Considering Its Degradation in Variable-Power Operations. *Appl. Energy* **2022**, *324*, 119760. [CrossRef]
28. Frensch, S.H.; Fouda-Onana, F.; Serre, G.; Thoby, D.; Araya, S.S.; Kær, S.K. Influence of the Operation Mode on PEM Water Electrolysis Degradation. *Int. J. Hydrogen Energy* **2019**, *44*, 29889–29899. [CrossRef]
29. Papakonstantinou, G.; Algara-Siller, G.; Teschner, D.; Vidaković-Koch, T.; Schlögl, R.; Sundmacher, K. Degradation Study of a Proton Exchange Membrane Water Electrolyzer under Dynamic Operation Conditions. *Appl. Energy* **2020**, *280*, 115911. [CrossRef]
30. Schofield, L.; Paren, B.; Macdonald, R.; Shao-Horn, Y.; Mallapragada, D. Dynamic Optimization of Proton Exchange Membrane Water Electrolyzers Considering Usage-Based Degradation. *AIChE J.* **2025**, *71*, e18635. [CrossRef]
31. Lai, F.; Budaev, S.; Hundven, I.K.; Balseiro, P.; Handeland, S.O.; Rønnestad, I. Influence of Water Temperature on Feed Intake, Appetite Control, and Energy Allocation in Atlantic Salmon (*Salmo salar*) Post-Smolt. *Front. Physiol.* **2025**, *16*, 1646208. [CrossRef] [PubMed]
32. Fujimoto, S.; van Hoof, H.; Meger, D. Addressing Function Approximation Error in Actor-Critic Methods. In Proceedings of the 35th International Conference on Machine Learning (ICML), Stockholm, Sweden, 10–15 July 2018; pp. 1587–1596.
33. Jonkman, B.; Platt, A.; Branlard, E.; Sprague, M.; Ross, H.; Hayman, G.; Vijayakumar, G.; Buckley, C.; Damiani, R.; des Lauriers, A.; et al. *OpenFAST User's Guide and Theory Manual (Version 4.2.1)*; National Renewable Energy Laboratory: Golden, CO, USA, 2025. Available online: https://openfast.readthedocs.io (accessed on 24 September 2026).
34. Gaertner, E.; Rinker, J.; Sethuraman, L.; Zahle, F.; Anderson, B.; Barter, G.; Abbas, N.; Meng, F.; Bortolotti, P.; Skrzypinski, W.; et al. *Definition of the IEA 15-Megawatt Offshore Reference Wind Turbine*; NREL Technical Report NREL/TP-5000-75698; National Renewable Energy Laboratory: Golden, CO, USA, 2020. [CrossRef]
35. Allen, C.; Viselli, A.; Dagher, H.; Goupee, A.; Gaertner, E.; Abbas, N.; Hall, M.; Barter, G. *Definition of the UMaine VolturnUS-S Reference Platform Developed for the IEA Wind 15-Megawatt Offshore Reference Wind Turbine*; NREL Technical Report NREL/TP-5000-76773; National Renewable Energy Laboratory: Golden, CO, USA, 2020. [CrossRef] [PubMed]
36. Abbas, N.J.; Zalkind, D.S.; Pao, L.; Wright, A. A Reference Open-Source Controller for Fixed and Floating Offshore Wind Turbines. *Wind Energy Sci.* **2022**, *7*, 53–73. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.