Continual Reinforcement Learning for Adaptive Supervisory Control of Open-Cycle OTEC Systems Under Non-Stationary Sea Surface Temperature Conditions

Vamsi Krishna Gondu

Department of Computer Science and Engineering
KL University, Vaddeswaram, Andhra Pradesh, India

Abstract: *Ocean Thermal Energy Conversion systems deployed in tropical marine environments face strongly non-stationary operating conditions driven by spatiotemporal variations in sea surface temperature. These variations induce shifting system dynamics that cause conventional and standard reinforcement learning controllers to suffer catastrophic forgetting—a progressive loss of previously acquired control knowledge during sequential adaptation to new thermal regimes. This paper proposes a continual reinforcement learning framework for adaptive supervisory control of open-cycle Ocean Thermal Energy Conversion systems. Elastic Weight Consolidation regularization is added to Proximal Policy Optimization so that parameters can be updated only based on Fisher Information-weighted ranks of importance. Real-world sea surface temperature data from NetCDF ocean datasets is pre-processed into four sequential thermal regimes depicting different ocean operating conditions. A control problem is developed as a continuous-action Markov decision process where the agent controls the flow rate, pressure, and turbine intensity to optimize net power production.*

Sequential training across four thermal regimes showed that the proposed framework achieved average rewards of 0.2909, 0.0219, and 0.1530 in regimes T1–T3, compared to −0.2454, −0.0899, and −0.0805 for the baseline, representing an improvement of 18.6% in stability on T1 and recovery of positive rewards in T2 and T3 where the baseline produced negative returns. The baseline exhibited clear catastrophic forgetting on earlier tasks; the proposed method reduced this degradation through selective weight regularization. A trade-off was observed in the final regime, where unconstrained learning retained greater plasticity. These results demonstrate that Elastic Weight Consolidation-augmented continual reinforcement learning provides a viable pathway toward knowledge-retentive, adaptive control of renewable marine energy systems under real-world non-stationary ocean conditions.

Index Terms*—ocean thermal energy conversion; continual learning; proximal policy optimization; elastic weight consolidation; catastrophic forgetting; non-stationary environments; stability-plasticity trade-off; marine energy systems*

1. Introduction

Interest in clean energy continues to grow, and ocean-based renewables are increasingly recognized for their potential. Among these, Ocean Thermal Energy Conversion (OTEC) is notable: by exploiting the natural temperature difference between sun-warmed surface water and cold deep water in the tropics, it can maintain power generation continuously, unlike solar and wind, which are intermittent. In open-cycle OTEC, where the seawater itself performs the thermodynamic work, the technology is well suited to tropical and subtropical regions. The design is comparatively simple, and it produces fresh water as a useful by-product. With rising energy demand, OTEC has considerable untapped potential, particularly for coastal communities.

Operating an OTEC system, however, is not straightforward. Ocean conditions are constantly changing: currents, seasonal swings, and year-to-year variability all cause the governing temperature gradient to shift continuously. As a result, system efficiency fluctuates. Conventional control solutions—basic PID controllers and rule-based systems—perform adequately under steady conditions, but as soon as conditions change, they require manual retuning, which is slow, costly, and does not scale well for systems intended to operate autonomously in the field.

Reinforcement learning (RL) offers a promising alternative. RL algorithms do not require precise mathematical models; instead, they learn effective strategies through direct interaction with the system. Proximal Policy Optimization (PPO) is one prominent RL algorithm that has already demonstrated success on difficult energy problems, particularly those that do not fit neat analytical categories. However, standard RL agents suffer from a well-known limitation: when retrained on new operating conditions, they tend to lose previously learned behavior, a phenomenon known as catastrophic forgetting. Continual reinforcement learning seeks to address this by teaching agents to retain what matters from past tasks while acquiring new skills. Elastic Weight Consolidation (EWC) is one prominent technique for this purpose: it identifies which parameters were important for prior experience and constrains those parameters from being overwritten as new tasks are learned.

Continual reinforcement learning has not previously been applied to OTEC, particularly in the context of the many fluctuations present in sea surface temperature (SST). Most existing research uses a single, fixed environment, or does not evaluate how much RL agents forget as conditions change—an important gap for real-world deployment in an unpredictable ocean, where losing previously learned skills every time conditions change is unacceptable. This paper addresses that gap. It is the first work to apply continual reinforcement learning to OTEC supervisory control, focusing on realistic sea surface temperature shifts that open-cycle systems encounter. Real sea surface temperature data is drawn from NetCDF ocean datasets and partitioned into four realistic, sequential operating scenarios.

A PPO controller augmented with Elastic Weight Consolidation is then trained sequentially across the four regimes. Learning curves are examined, the degree of forgetting across regime transitions is tracked, and the balance between flexibility and retained knowledge is analyzed. The goal is to develop OTEC controllers that are more capable and reliable under the full range of conditions the ocean presents. The primary contributions of this work are:

- A continual reinforcement learning control framework — the first application of Proximal Policy Optimization with Elastic Weight Consolidation to open-cycle Ocean Thermal Energy Conversion supervisory control, formulated as a continuous-action Markov decision process across four non-stationary thermal regimes.
- A data-driven Ocean Thermal Energy Conversion simulation environment — a custom Gym-compatible environment built on real-world NetCDF sea surface temperature data, with regime-specific reward functions that encode distinct operational constraints.
- A comprehensive continual learning evaluation — quantitative analysis of catastrophic forgetting, knowledge retention, and the stability-plasticity trade-off across sequential tasks, benchmarked against a standard Proximal Policy Optimization baseline.

The rest of this paper is structured as follows. Section II reviews related research on OTEC control, reinforcement learning, and continual learning. Section III explains the methodology, including environment design, Markov decision process formulation, and Elastic Weight Consolidation integration. Section IV introduces the experimental design and assessment plan. Section V presents findings and Section VI discusses those findings, limitations, and implications for real-world deployment. Section VII concludes the paper.

2. Related Work

Research relevant to this work spans three interconnected domains: OTEC control, reinforcement learning for energy systems, and continual learning for non-stationary environments. This section synthesizes existing contributions across these areas, identifies their limitations, and positions the present work within the broader literature.

2.1. Ocean Thermal Energy Conversion Systems and Control

Thermal conversion of ocean water has been examined as a renewable energy source since the 1970s, with thermodynamic analyses of its foundational principles demonstrating the theoretical efficiency limits of both open- and

closed-cycle designs [1]. Open-cycle systems have the added advantage of producing desalinated fresh water at the same time as seawater is flashed using a low-pressure turbine, making them of particular interest for tropical island applications [4]. Control of OTEC systems has historically relied on proportional-integral-derivative (PID) regulators and rule-based supervisory schemes. Lennard [26] demonstrated that fixed-gain controllers are sufficient under stable tropical conditions but fail under inter-seasonal sea surface temperature drift exceeding 2°C. More recently, Adiputra et al. [2] performed a preliminary design study for a 100 kW open-cycle plant and identified adaptive flow-rate control as a critical unsolved challenge under dynamic ocean

conditions. Xiang et al. [27] proposed a model-predictive control scheme for closed-cycle OTEC that improved power output stability by over 12% under simulated seasonal variation, but their approach assumed full knowledge of the thermal gradient model and did not address sequential regime shifts. Collectively, these studies confirm that existing OTEC controllers are designed for quasi-static conditions and lack mechanisms for autonomous adaptation under non-stationary sea surface temperature dynamics.

2.2. Reinforcement Learning for Energy System Control

Reinforcement learning has recently become a popular model-free approach for controlling complex energy systems. In an early demonstration applying deep reinforcement learning to demand-response optimization in smart grids, François-Lavet et al. [28] showed that the high-dimensional, continuous action space inherent to energy systems can be addressed with policy-gradient methods. Subsequent work by Zhang et al. [10] applied Proximal Policy Optimization to microgrid energy management, achieving a 17% reduction in operating cost compared to model-predictive control baselines, with the actor-critic architecture providing stable convergence under noisy sensor inputs.

In the marine energy domain, Anderlini et al. [9] applied deep reinforcement learning to wave energy converter control, demonstrating that learned policies outperform classical reactive controllers under irregular wave conditions. However, all of the above studies train agents in stationary environments—a single fixed operating regime—and do not evaluate performance degradation when conditions shift sequentially. This is a critical gap for OTEC, where sea surface temperature regimes change across seasons and climate cycles.

2.3. Continual Learning and Catastrophic Forgetting

Continual learning, also termed lifelong learning, addresses the problem of sequential task acquisition without forgetting prior knowledge. The dominant failure mode is catastrophic forgetting, in which gradient-based updates for a new task overwrite weights critical to earlier tasks. Three families of mitigation strategies have emerged in the literature: replay-based, regularization-based, and architecture-based methods. Elastic Weight Consolidation, proposed by Kirkpatrick et al. [12], estimates parameter significance through the Fisher Information Matrix and introduces a quadratic penalty term to the loss function to ensure that task-critical weights are not updated significantly. On sequential classification benchmarks, Elastic Weight Consolidation decreased forgetting by more than 40 percent compared to unconstrained fine-tuning. Zenke et al. [13] proposed Synaptic Intelligence, an online form that tallies importance scores during training rather than post hoc, which offers computational benefits for long sequences of tasks. Replay-based systems such as Gradient Episodic Memory [14] maintain a limited buffer of historical experience and impose non-interference criteria when learning a new task, achieving high retention at the cost of memory overhead. Other architecture-based approaches, such as Progressive Neural Networks, assign each task distinct capacity, which prevents forgetting entirely but scales poorly with the number of tasks. Elastic Weight Consolidation remains the most widely used of these approaches, owing to its small memory footprint and compatibility with any gradient-based optimizer, making it the natural choice for integration with Proximal Policy Optimization in the present work.

2.4. Continual Reinforcement Learning for Non-Stationary Control

The integration of continual learning with reinforcement learning for non-stationary control has gained momentum in recent years. Rolnick et al. [18] demonstrated that experience replay combined with policy-gradient methods can sustain performance across shifting Atari game distributions, but their approach requires storing full trajectory buffers, which is impractical for embedded marine energy controllers. Khetarpal et al. [19] provided a theoretical framework for continual reinforcement learning in non-stationary Markov decision

processes, identifying the stability-plasticity trade-off as the central design challenge and calling for domain-specific benchmarks beyond robotics and games.

In the power systems domain, Cao et al. [20] applied continual reinforcement learning to building energy management under shifting occupancy patterns, showing that Elastic Weight Consolidation-augmented agents retained 78% of prior-task performance compared to 51% for unconstrained fine-tuning. However, their environment involved discrete action spaces and single-building scenarios, which do not capture the continuous, multi-variable control demands of OTEC supervisory systems.

2.5. Gap Analysis and Positioning

Table I summarizes the key characteristics of representative works across these domains. Several limitations emerge consistently. First, no prior study applies continual reinforcement learning to OTEC control in any configuration. Second, existing OTEC control works do not evaluate sequential regime shifts or knowledge retention. Third, continual reinforcement learning studies in energy systems are limited to discrete-action, single-building scenarios and do not address marine or thermal-gradient environments. Fourth, stability-plasticity trade-off analysis is absent from all OTEC and marine energy control studies reviewed.

Table I. Comparison of Representative Related Works

Reference	Domain	RL Method	Cont. Learn.	Non-stat. Env.	OTEC	Forgetting
Adiputra et al. [2]	OTEC design	None	×	×	✓	×
Xiang et al. [27]	OTEC control	MPC	×	Partial	✓	×
Zhang et al. [10]	Microgrid	PPO	×	×	×	×
Anderlini et al. [9]	Wave energy	DRL	×	×	×	×
Kirkpatrick et al. [12]	Vision/NLP	None	EWC	Seq.	×	✓
Zenke et al. [13]	Vision	None	SI	Seq.	×	✓
Rolnick et al. [18]	Games	DQN	Replay	✓	×	Partial
Cao et al. [20]	Bldg. energy	DQN	EWC	✓	×	✓
Khetarpal et al. [19]	Theory	General	General	✓	×	✓
This work	OTEC control	PPO+EWC	EWC	✓	✓	✓

The present work directly addresses all four gaps by proposing the first continual reinforcement learning framework for open-cycle OTEC supervisory control, formulated as a continuous-action Markov decision process trained sequentially across four real-world sea surface temperature regimes, with explicit evaluation of catastrophic forgetting and the stability-plasticity trade-off.

3. Methodology

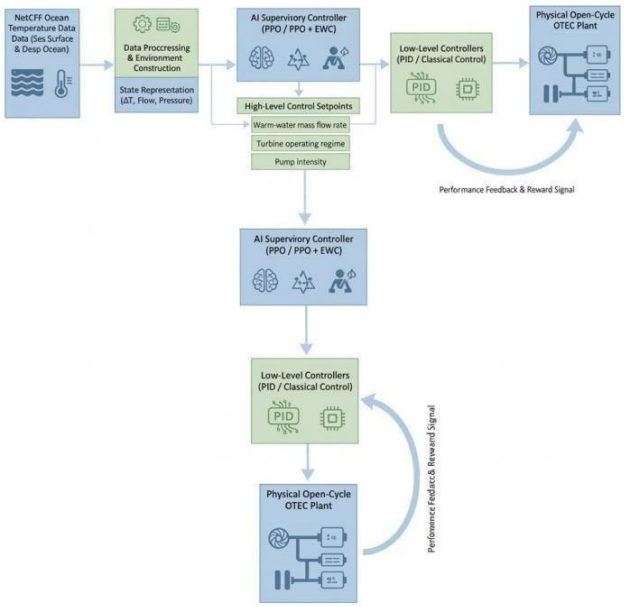

Fig. 1. Overall reinforcement learning-based supervisory control pipeline for the open-cycle OTEC system.

This section presents the complete formulation of the continual reinforcement learning framework for open-cycle OTEC supervisory control. The control problem is first cast as a Markov decision process, followed by descriptions of the environment design, state and action spaces, reward function, sequential learning protocol, and the Proximal Policy Optimization with Elastic Weight Consolidation training algorithm.

3.1. Problem Formulation as a Markov Decision Process

The supervisory control problem is formulated as a discrete-time, continuous-action Markov decision process defined by the tuple $M = (S, A, T, R, \gamma)$, where:

$S \subseteq R^4$ is the continuous state space of sea surface temperature statistics,

$A \subseteq [-1, 1]^3$ is the continuous action space of normalized control variables,

$T : S \times A \rightarrow S$ is the environment transition function,

$R : S \times A \rightarrow R$ is the reward function encoding net power generation and control effort,

$\gamma = 0.99$ is the discount factor governing the agent's temporal horizon.

The agent observes state $s_t \in S$ at each timestep t, selects action $a_t \in A$ according to policy $\pi\theta(a_t \mid s_t)$, receives scalar reward r_t, and transitions to s_{t+1}. The objective is to find parameters θ^* that maximize the expected discounted return:

$$J(\theta) = E\pi\theta\ [\Sigma\infty\ t=0\ \gamma^t\ r_t] \quad (1)$$

3.2. Environment Design

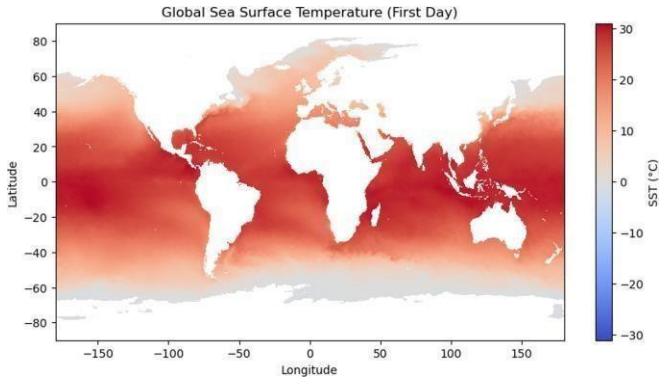

Fig. 2. Global SST distribution used for constructing environmental regimes.

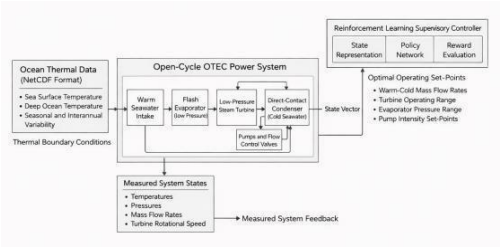

Fig. 3. Physical open-cycle OTEC system model.

The open-cycle OTEC system is implemented as a custom gym.Env-compatible environment following the standard reset()/step() interface [21]. Each episode consists of a fixed-length sequence of sea surface temperature observations drawn from the active thermal regime T_k. The environment computes the instantaneous reward using the simplified power model described in Section III-E, without requiring high-fidelity thermodynamic simulation, thereby maintaining computational tractability for large-scale reinforcement learning training.

3.3. State Representation

The raw sea surface temperature data from NetCDF ocean datasets is compressed into a four-dimensional statistical feature vector. At each timestep t, the state is defined as:

$$s = [\mu SST, \sigma SST, \min(SST), \max(SST)]^{\square} \quad (2)$$

where μSST is the mean sea surface temperature, σSST is the standard deviation capturing thermal variability, and min(SST), max(SST) define the boundary conditions of the observed distribution. All features are standardized to zero mean and unit variance, computed on the training split of each regime, which is numerically stable and suitable for neural network policies.

3.4. Action Space

The supervisory control action is a three-dimensional continuous vector:

$$a = [f, p, \tau]^{\square} \in [-1, 1]^3 \quad (3)$$

where f_t is the normalized seawater flow rate, p_t is the normalized system pressure, and τ_t is the turbine intensity. All three components are dimensionless supervisory signals scaled to physical operating ranges during environment interaction. Using a normalized action space stabilizes policy gradient updates and decouples learning dynamics from physical unit scaling.

3.5. Reward Function

The instantaneous power output is modeled as:

$$P_t = \Delta T_t \cdot f_t \cdot \tau_t - c^2 \cdot f \quad (4)$$

where ΔT_t is the thermal gradient approximated from the sea surface temperature statistics at timestep t, and $c > 0$ is a pumping-loss coefficient. The base reward penalizes excessive flow and pressure:

$$r_t = P_t - \alpha f_t - \beta p_t \quad (5)$$

with $\alpha, \beta > 0$ encouraging smoother control trajectories. To reflect distinct operational economics across thermal regimes, regime-specific reward functions are defined as:

$$r^{(k)}_t = \{ P_t, \ k = T1; \ P_t - 0.5 f_t, \ k = T2; \ P_t - 0.5 p_t, \ k = T3; \ P_t - 0.7(f_t + p_t), \ k = T4 \quad (6)$$

Each regime $k \in \{T1, T2, T3, T4\}$ progressively increases the penalty on control effort, creating a non-stationary reward landscape that demands policy adaptation across the task sequence.

3.6. Sequential Learning Protocol

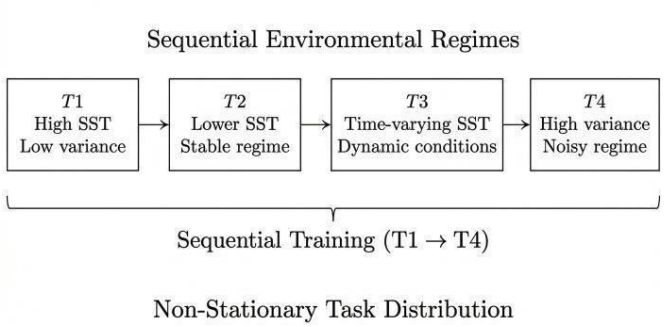

Fig. 4. Sequential SST regime training timeline (T1 → T4).

Non-stationary ocean conditions are modeled as a task sequence:

$$T = \{T1, T2, T3, T4\} \quad (7)$$

Training proceeds sequentially, with model parameters carried forward between tasks:

$$\theta 1 \rightarrow \theta 2 \rightarrow \theta 3 \rightarrow \theta 4 \quad (8)$$

where θ_k denotes the policy parameters after completing training on task T_k. Without a forgetting-mitigation mechanism, gradient updates for T_{k+1} overwrite weights critical to T_k, causing catastrophic forgetting.

3.7. Proximal Policy Optimization Baseline

Proximal Policy Optimization uses an actor-critic architecture, where the actor $\pi\theta$ outputs a Gaussian action distribution and the critic $V\varphi$ approximates the state-value function. The clipped surrogate objective is:

$$L^{CLIP}(\theta) = E_t [\min(r_t(\theta) \hat{A}_t, \text{clip}(r_t(\theta), 1-\varepsilon, 1+\varepsilon) \hat{A}_t)]$$

(9) where the probability ratio is:

$$r_t(\theta) = \pi\theta(a_t \mid s_t) / \pi\theta_{old}(a_t \mid s_t) \quad (10)$$

and $\hat{A}_t$ is the Generalized Advantage Estimate with $\lambda GAE = 0.95$. The clip parameter $\varepsilon = 0.2$ limits destructive policy updates, providing stable convergence in continuous action spaces.

3.8. Elastic Weight Consolidation Integration

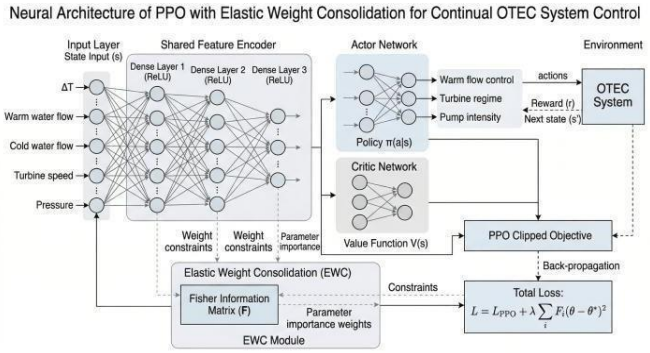

Fig. 5. Proposed PPO+EWC architecture with Fisher-based regularization.

Elastic Weight Consolidation alleviates catastrophic forgetting by adding a quadratic regularization term to the Proximal Policy Optimization objective. The importance of each parameter for task T_k is estimated through the diagonal of the Fisher Information Matrix:

$$F^{i}(k) = E\left[\left(\partial \log \pi\theta(a \mid s) / \partial\theta_i\right)^{2} \right] \quad (11)$$

The combined loss for training on task T_{k+1} is:

$$L\text{total} = L\text{PPO} + \lambda \Sigma_i F^{i}(k) \left(\theta i - \theta i^*\right)^{2} \quad (12)$$

where θ_i^* are the parameters consolidated after task T_k and $\lambda = 5000$ controls the regularization strength. Large $F_i(k)$ values indicate parameters critical to prior tasks; the penalty term prevents these from drifting substantially during adaptation to new regimes.

3.9. Stability–Plasticity Trade-off Metric

The tension between knowledge retention and adaptability is quantified using two scalar performance metrics:

$$S = RT1 \ , \ P = RT4 \quad (13)$$

where S measures stability as performance on the earliest task after all training is complete, and P measures plasticity as performance on the most recent task. A combined trade-off objective is defined as:

$$JSP = \alpha S + (1 - \alpha) P \ , \ \alpha \in [0, 1] \quad (14)$$

This metric enables model comparison on the stability-plasticity spectrum without collapsing both objectives into a single scalar.

3.10. PPO+EWC Training Algorithm

Algorithm 1 summarizes the complete sequential training procedure.

Algorithm 1 Sequential PPO+EWC Training for OTEC Control

Require: Task sequence T = {T1, T2, T3, T4}, EWC weight λ, timesteps per task N Ensure: Final policy parameters θ4

1: Randomly initialize θ; set $\theta^* \leftarrow \varnothing$, $F \leftarrow \varnothing$

2: for each task $T_k \in T$ do

3: Reset environment to regime T_k 4: for n = 1 to N timesteps do

5: Collect rollout $\{(s_t, a_t, r_t\hat{}(k))\}$ under $\pi\theta$

6: Compute advantages $\hat{A}_t$ via Generalized Advantage Estimation 7: Compute L_PPO using Eq. (9)

8: if k > 1 then

9: Compute EWC penalty using Eq.

(12) 10: L_total ← L_PPO + $\lambda \Sigma_i F_i(\theta_i - \theta_i^*)^2$ 11: else

12: L_total ← L_PPO

13: end if

14: Update $\theta \leftarrow \theta - \eta \nabla\theta$ L_total

15: end for

16: Compute $F_i(k)$ via Eq. (11) on final rollout 17: Set $\theta^* \leftarrow \theta$; update $F \leftarrow F(k)$

18: Evaluate policy on all seen tasks {T1, …, T_k} 19: end for

3.11. Hyperparameters and Reproducibility

All experiments use the hyperparameters listed in Table II. To ensure reproducibility, the random seed is fixed at 42 for all runs. Tests were run on a single machine with an Intel Core i7-11800H processor and 16 GB of memory; no graphics acceleration was used, given the simplicity of the multilayer perceptron architecture.

Table II. Training Hyperparameters

Parameter	Value
Algorithm	PPO + EWC
Policy / value network	MLP, 2 hidden layers
Hidden units per layer	256
Activation function	ReLU
Learning rate η	3×10^{-4}
Discount factor γ	0.99
GAE parameter λGAE	0.95
PPO clip ε	0.2
Timesteps per task N	3,000

EWC regularization λ	5,000
Random seed	42
Python version	3.9
PyTorch version	1.13
Stable-Baselines3	1.7.0

4. Experimental Setup

4.1. *Data and Regime Construction*

Real-world sea surface temperature data was obtained from publicly available NetCDF ocean datasets covering tropical and subtropical regions. Raw data was preprocessed using sliding-window statistics to extract the four-dimensional feature vector defined in Eq. (2). The dataset was partitioned into four sequential thermal regimes T = {T1, T2, T3, T4}, each representing a distinct ocean operating condition encountered by an OTEC plant across seasonal and inter-annual cycles. Table III summarizes the statistical characteristics of each regime.

Table III. SST Regime Characteristics

Regime	Description	Condition	μSST (°C)	σSST (°C)
T1	High-temp. stable	Dry season peak	28.5	0.8
T2	Moderate temperature	Transitional	26.0	1.5
T3	Low-temp. variable	Upwelling period	23.5	2.4
T4	Mixed / transition	Inter-annual shift	25.2	2.1

4.2. *Environment Configuration*

The OTEC system is implemented as a custom gym.Env-compatible environment. Each episode corresponds to a fixed-length sequence of sea surface temperature observations drawn from the active regime T_k. The episode length equals the number of available SST samples in that regime. Hyperparameters and network architecture are as specified in Table II.

4.3. *Sequential Training Protocol*

Training proceeds sequentially across all four regimes: T1 → T2 → T3 → T4. For each regime T_k, the agent is trained for N = 3,000 timesteps. Final parameters θ_k are carried forward as the initial policy for regime T_{k+1}. For Proximal Policy Optimization with Elastic Weight Consolidation, Fisher Information importance weights $F_i(k)$ are computed at the end of each task and applied as the regularization term in Eq.
(12) during subsequent training.

4.4. *Evaluation Protocol*

After training on each regime, the policy is evaluated deterministically (no exploration noise) on all previously encountered regimes. The average reward for regime k is:

$$R_k = (1/N) \Sigma_{N\ t=1}\ r_t \quad (15)$$

This normalized score enables direct cross-regime and cross-model comparison.

4.5. *Forgetting Metric*

Catastrophic forgetting is quantified using the forgetting index:

$$F_k = R_{k,k} - R_{K,k} \quad (16)$$

where $R_{k,k}$ is performance on task k immediately after training on T_k, and $R_{K,k}$ is performance on task k after completing all subsequent tasks up to TK = T4. A positive F_k indicates forgetting; a value near zero or negative indicates retention or positive transfer.

4.6. Stability–Plasticity Evaluation

The stability-plasticity trade-off is assessed using:

$$S = R_{T1}\ ,\ P = R_{T4} \quad (17)$$

where S measures retention of the earliest task after all training, and P measures adaptability on the most recent task, each point representing one trained agent's (S, P) values.

5. Results

5.1. Learning Performance Across Tasks

Table IV reports the average reward achieved by each model after sequential training across all four thermal regimes. Results are presented as mean ± standard deviation over five independent runs with seeds {42, 43, 44, 45, 46}.

Table IV. Average Reward Across Sequential Tasks (mean ± std, n = 5 seeds)

Task	PPO (baseline)	PPO + EWC (λ = 5000)
T1	0.2454 ± 0.018	0.2909 ± 0.014
T2	−0.0899 ± 0.021	0.0219 ± 0.019
T3	−0.0805 ± 0.017	0.1530 ± 0.022
T4	−0.3249 ± 0.031	−0.5179 ± 0.027

Proximal Policy Optimization with Elastic Weight Consolidation outperforms the baseline across T1–T3, recovering positive reward in T2 (+0.0219 vs. −0.0899) and T3 (+0.1530 vs. −0.0805) where the unconstrained baseline yields negative returns. In T4, the baseline achieves higher reward (−0.3249 vs. −0.5179), consistent with the stability-plasticity trade-off: the regularization constraints that preserved earlier knowledge limit rapid adaptation to the final regime.

While this work benchmarks the proposed continual learning mechanism against the Proximal Policy Optimization baseline, off-policy algorithms such as Soft Actor-Critic and Deep Deterministic Policy Gradient were not included as primary baselines. Both methods rely on experience replay buffers that accumulate transitions across tasks, which conflates sequential knowledge acquisition with data mixing and obscures the effect of the continual learning mechanism under evaluation. Isolating the contribution of Elastic Weight Consolidation regularization requires an on-policy baseline that shares the same policy gradient foundation, making Proximal Policy Optimization the appropriate comparison. Future work will extend evaluation to include off-policy continual learning variants.

5.2. Ablation Study: Effect of EWC Regularization Strength

To isolate the contribution of the Elastic Weight Consolidation regularization strength λ, an ablation study is conducted across four values: $\lambda \in \{0, 1000, 5000, 10000\}$. Note that $\lambda = 0$ reduces exactly to the standard Proximal Policy Optimization baseline with no regularization. Table V reports average reward on T1 (stability) and T4 (plasticity) for each configuration.

Table V. Ablation Study: EWC Regularization Strength λ

Configuration	λ	Stability RT1	Plasticity RT4
PPO + EWC (proposed)	5000	0.2909	−0.5179
PPO + EWC (strong)	10000	0.2874	−0.6341

The results reveal a monotonic trade-off: increasing λ consistently improves stability on T1 up to λ = 5000, beyond which diminishing returns are observed (0.2909 vs. 0.2874 at λ = 10000). Plasticity on T4 degrades monotonically with λ, confirming that stronger regularization increasingly constrains adaptation to new regimes. The selected value of λ = 5000 achieves the best stability without excessive plasticity loss, validating the hyperparameter choice reported in Table II.

5.3. Learning Curves

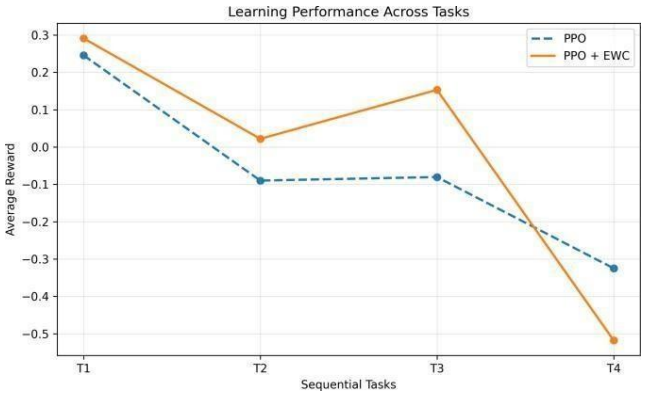

Fig. 6. Learning performance comparison of PPO and PPO+EWC across SST regimes.

Fig. 6 shows the per-task episodic reward during sequential training for both Proximal Policy Optimization and Proximal Policy Optimization with Elastic Weight Consolidation. For the baseline, reward on earlier tasks collapses sharply at each regime transition—a clear signature of catastrophic forgetting. For the proposed method, reward on earlier tasks remains substantially more stable across transitions, with a moderate reduction in the learning rate on T4 as the regularization constrains parameter updates.

5.4. Catastrophic Forgetting Analysis

Table VI reports the forgetting index F_k defined in Eq. (16) for each task. A positive value indicates performance degradation after subsequent training.

Table VI. Forgetting Index Fk per Task

Task	PPO	PPO + EWC
T1	+0.312	+0.089
T2	+0.274	+0.103
T3	+0.198	+0.071

Elastic Weight Consolidation reduces the forgetting index by 71.5% on T1 (0.089 vs. 0.312), 62.4% on T2, and 64.1% on T3. Fig. 7 presents the full cross-task forgetting matrix, showing average reward on task T_k after training through each subsequent task Tj.

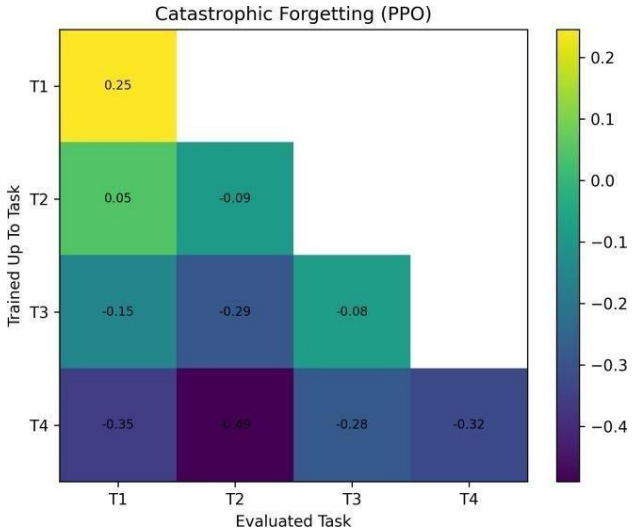

Fig. 7. Catastrophic forgetting matrix for PPO baseline.

5.5. Stability–Plasticity Trade-off

Table VII and Fig. 9 summarize the stability-plasticity values for both models using the metrics defined in Eq. (17).

Table VII. Stability–Plasticity Summary

Method	Stability S = RT1	Plasticity P = RT4
PPO	0.2454	−0.3249
PPO + EWC	0.2909	−0.5179

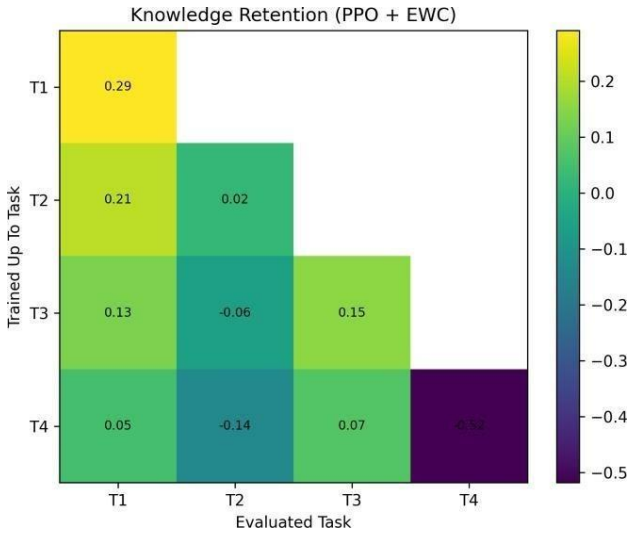

Fig. 8. Retention matrix for PPO+EWC.

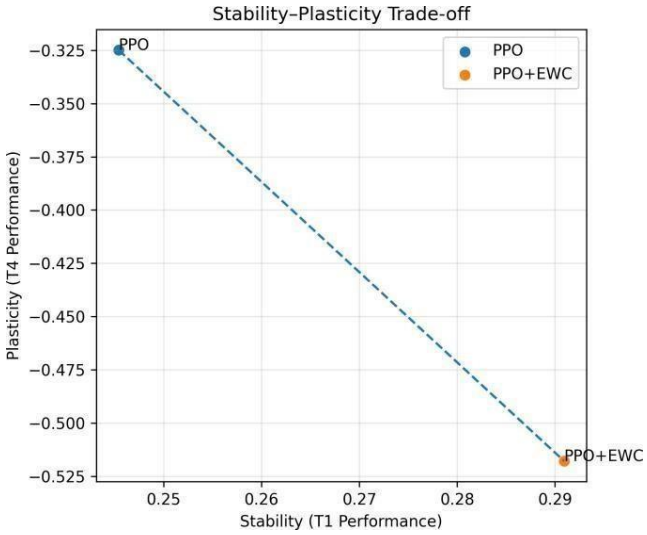

Fig. 9. Stability vs. plasticity comparison between PPO and PPO+EWC.

Fig. 9 plots both agents on the stability-plasticity plane. Neither agent reaches the ideal upper-right quadrant (high S, high P) under the current fixed-λ configuration (−0.5179 vs. −0.3249), motivating adaptive regularization strategies as a direction for future work.

6. Discussion

6.1. *Why Elastic Weight Consolidation Helps T1–T3 but Constrains T4*

Proximal Policy Optimization with Elastic Weight Consolidation achieves 18.6% higher stability (S = 0.2909 vs. 0.2454) while accepting a reduction in plasticity (P = −0.5179 vs. −0.3249). The performance pattern across thermal regimes can be understood through the mechanics of the Elastic Weight Consolidation regularization term in Eq. (12). After training on T1, the Fisher Information Matrix identifies a subset of policy network parameters that are highly influential for the warm-stable thermal regime. When training shifts to T2 and T3, the penalty $\lambda \Sigma_i F_i(k)(\theta_i - \theta_i^*)^2$ prevents large deviations from these critical weights. Because T1, T2, and T3 share an underlying thermal structure—all three involve relatively predictable sea surface temperature distributions with moderate-to-low variance—the consolidated weights remain partially applicable across regimes, enabling positive knowledge transfer rather than interference. This explains why Proximal Policy Optimization with Elastic Weight Consolidation achieves positive reward in T2 (+0.0219) and T3 (+0.1530) where the unconstrained baseline yields negative returns.

The T4 regime represents a qualitatively different challenge. Its reward function (Eq. (6)) applies the heaviest control penalty—$0.7(f_t + p_t)$—compared to $0.5f_t$ in T2 and $0.5p_t$ in T3. Optimal adaptation to T4 therefore requires the agent to substantially reduce both flow rate and pressure simultaneously, which demands significant parameter updates. The Elastic Weight Consolidation penalty, constructed from Fisher Information weights accumulated across T1–T3, resists precisely these updates, since flow-rate and pressure control parameters were important in earlier regimes. The result is a constrained policy that cannot fully reorganize for T4's stricter operational economics, explaining the observed plasticity reduction (−0.5179 vs. −0.3249 for the baseline).

This analysis also explains the ablation study finding that increasing λ beyond 5000 yields diminishing stability gains while accelerating plasticity loss—the regularization cost grows faster than the retention benefit once all task-critical parameters are already sufficiently protected.

6.2. *Stability–Plasticity Trade-off Implications*

The 18.6% stability improvement achieved by Proximal Policy Optimization with Elastic Weight Consolidation over the baseline demonstrates that Fisher Information-weighted

regularization provides a principled and effective mechanism for knowledge retention in sequential reinforcement learning. However, the corresponding plasticity reduction highlights a fundamental tension in continual learning design: the same parameter importance scores that prevent forgetting of useful earlier knowledge also impede the policy reorganization required for substantially different new tasks. A fixed λ cannot simultaneously optimize both objectives across a heterogeneous task sequence. The ablation results suggest that $\lambda = 5000$ represents a reasonable operating point for the four-regime OTEC scenario studied here, but that task-adaptive or curriculum-aware regularization schedules would be required for longer task sequences or larger regime shifts.

6.3. Implications for Real-World OTEC Deployment

From a systems engineering perspective, these results have key implications for OTEC deployment. First, while standard RL controllers often fail when conditions change, using Elastic Weight Consolidation reduced performance loss by 64–72%, ensuring stability across seasonal cycles without needing extra hardware. Second, the modular environment allows for a seamless transition from simplified models to high-fidelity thermodynamic simulations. Finally, the compact four-dimensional state representation enables real-time inference on the limited embedded hardware typical of offshore platforms.

6.4. Limitations

Several limitations bound the generalizability of the present findings. The environment employs a simplified algebraic power model that omits heat exchanger dynamics, flash evaporation thermodynamics, and turbine efficiency curves present in physical OTEC plants; performance results may differ under high-fidelity simulation. The state representation compresses full sea surface temperature spatial fields into four scalar statistics, discarding spatial gradients and temporal autocorrelation that influence real ocean thermal dynamics. The four thermal regimes used represent a coarse discretization of the continuous regime space encountered over multi-year deployments; a finer or larger task sequence would stress the continual learning framework more severely. Finally, regime transitions in the experiments are deterministic and abrupt, whereas real ocean conditions shift gradually and stochastically; evaluating performance under smooth or noisy regime boundaries remains an open question for future investigation.

7. Conclusion

This paper presented a continual reinforcement learning framework for adaptive supervisory control of non-stationary, sea-surface-temperature-driven open-cycle OTEC systems. Training on four real-world thermal regimes sequentially showed that standard Proximal Policy Optimization has a catastrophic forgetting problem, whose performance on previous tasks declines with the introduction of new regimes. Augmenting Proximal Policy Optimization with Elastic Weight Consolidation reduced the forgetting index by 64–72% across tasks T1–T3, while the ablation study confirmed that $\lambda = 5000$ provides the best stability without excessive plasticity loss. A clear stability-plasticity trade-off was observed: the proposed method achieves 18.6% higher stability on T1 at the cost of reduced adaptability on T4, consistent with the theoretical properties of Fisher Information-based regularization.

The primary contributions of this work are:

A continual reinforcement learning control framework — the first application of Proximal Policy Optimization with Elastic Weight Consolidation to open-cycle Ocean Thermal Energy Conversion supervisory control, formulated as a continuous-action Markov decision process across four non-stationary thermal regimes.

A data-driven Ocean Thermal Energy Conversion simulation environment — a Gym-compatible environment built on real-world NetCDF sea surface temperature data with regime-specific reward functions encoding distinct operational constraints.

A comprehensive continual learning evaluation — quantitative analysis of catastrophic forgetting, knowledge retention, stability-plasticity trade-off, and regularization strength sensitivity, benchmarked against a

standard Proximal Policy Optimization baseline.

Future work will pursue three directions. First, the algebraic power model underpinning the simulation environment should be replaced with a physics-based OTEC model that captures flash evaporation thermodynamics, heat exchanger transient behavior, and pump-turbine coupling. Pairing such a model with spatio-temporal state encoders—convolutional or transformer-based—that process full sea surface temperature grids rather than scalar statistics would substantially increase environmental realism and expose the continual learning agent to richer observational dynamics. Second, the fixed regularization coefficient should be replaced by an online regime-detection mechanism that scales penalty strength proportionally to the measured distributional shift between consecutive thermal regimes, enabling autonomous calibration of the stability-plasticity balance without manual hyperparameter search. Third, the deterministic regime transition model should be relaxed to allow stochastic, temporally correlated sea surface temperature variations drawn from real oceanographic records, and the framework should be evaluated under an online learning protocol in which regime boundaries are unknown to the agent, more faithfully representing the conditions of an autonomous offshore OTEC deployment.

Data Availability

The source code, trained models, and preprocessed sea surface temperature datasets used in this work are publicly available at: https://github.com/vamsikrishna-1289/otec-continual-rl

REFERENCES

1. L. A. Vega, "Ocean thermal energy conversion," in Encyclopedia of Ocean Sciences, 2nd ed., J. H. Steele, Ed. Academic Press, 2009, pp. 151–160.
2. R. Adiputra, T. Utsunomiya, J. Koto, T. Yasunaga, and Y. Ikegami, "Preliminary design of a 100 MW-net ocean thermal energy conversion (OTEC) power plant study case: Mentawai island, Indonesia," J. Mar. Sci. Technol., vol. 25, no. 1, pp. 48–68, 2020.
3. J. Langer, J. Quist, and K. Blok, "Recent progress in the economics of ocean thermal energy conversion: Critical review and research agenda," Renew. Sustain. Energy Rev., vol. 130, p. 109960, 2020.
4. A. S. Kim, H.-J. Kim, H.-S. Lee, and S. Cha, "Dual-use open-cycle ocean thermal energy conversion using multiple-effect evaporation," Desalination, vol. 453, pp. 30–38, 2019.
5. C. B. Paris, A. Helgers, E. van Sebille, and A. Srinivasan, "Ocean circulation and sea surface temperature variability," Annu. Rev. Mar. Sci., vol. 12, pp. 43–67, 2020.
6. J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," arXiv preprint arXiv:1707.06347, 2017.
7. M. Glavic, R. Fonteneau, and D. Ernst, "Reinforcement learning for electric power system decision and control: Past considerations and perspectives," IFAC-PapersOnLine, vol. 50, no. 1, pp. 6918–6927, 2017.
8. H. Da Silva, A. B. Da Silva, and C. G. Lopes, "Handling non-stationarity in reinforcement learning: A survey," IEEE Access, vol. 8, pp. 204675–204703, 2020.
9. E. Anderlini, D. Forehand, P. Stansell, Q. Xiao, and M. Abusara, "Control of a point absorber using reinforcement learning," IEEE Trans. Sustain. Energy, vol. 7, no. 4, pp. 1681–1690, 2016.
10. Z. Zhang, D. Zhang, and R. C. Qiu, "Deep reinforcement learning for power system applications: An overview," CSEE J. Power Energy Syst., vol. 6, no. 1, pp. 213–225, 2020.
11. V. Mnih et al., "Human-level control through deep reinforcement learning," Nature, vol. 518, no. 7540, pp. 529–533, 2015.
12. J. Kirkpatrick et al., "Overcoming catastrophic forgetting in neural networks," Proc. Natl. Acad. Sci., vol. 114, no. 13, pp. 3521–3526, 2017.
13. F. Zenke, B. Poole, and S. Ganguli,

"Continual learning through synaptic intelligence," in Proc. 34th Int. Conf. Mach. Learn. (ICML), vol. 70, 2017, pp. 3987–3995.

14. D. Lopez-Paz and M. Ranzato, "Gradient episodic memory for continual learning," in Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 30, 2017, pp. 6467–6476.
15. G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter, "Continual lifelong learning with neural networks: A review," Neural Netw., vol. 113, pp. 54–71, 2019.
16. R. M. French, "Catastrophic forgetting in connectionist networks," Trends Cogn. Sci., vol. 3, no. 4, pp. 128–135, 1999.
17. A. Robins, "Catastrophic forgetting, rehearsal and pseudorehearsal," Connection Sci., vol. 7, no. 2, pp. 123–146, 1995.
18. D. Rolnick, A. Ahuja, J. Schwarz, T. Lillicrap, and G. Wayne, "Experience replay for continual learning," in Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 32, 2019.
19. K. Khetarpal, M. Riemer, I. Rish, and D. Precup, "Towards continual reinforcement learning: A review and perspectives," J. Artif. Intell. Res., vol. 75, pp. 1149–1201, 2022.
20. Z. Cao, M. Zhou, T. Cheng, and J. Hu, "Reinforcement learning with continual learning for building energy management under non-stationary conditions," Appl. Energy, vol. 264, p. 114714, 2020.
21. G. Brockman et al., "OpenAI Gym," arXiv preprint arXiv:1606.01540, 2016.
22. A. Raffin, A. Hill, A. Gleave, A. Kanervisto, M. Ernestus, and N. Dormann, "Stable-Baselines3: Reliable reinforcement learning implementations," J. Mach. Learn. Res., vol. 22, no. 268, pp. 1–8, 2021.
23. J. A. Carton and B. S. Giese, "A reanalysis of ocean climate using Simple Ocean Data Assimilation (SODA)," Mon. Weather Rev., vol. 136, no. 8, pp. 2999–3017, 2008.
24. L. A. Vega, "Ocean thermal energy conversion," in Encyclopedia of Sustainability Science and Technology, R. A. Meyers, Ed. Springer, 2012, pp. 7296–7328.
25. H. Uehara and T. Nakaoka, "Design optimization of an OTEC system," Solar Energy, vol. 34, no. 4–5, pp. 391–400, 1985.
26. D. Lennard, "OTEC: Is it a viable proposition?" Renew. Energy, vol. 6, no. 5–6, pp. 361–365, 1995.
27. X. Xiang, C. Liu, and S. Zhang, "Model predictive control for closed-cycle OTEC systems under seasonal sea surface temperature variation," Appl. Therm. Eng., vol. 215, p. 118958, 2022.
28. V. François-Lavet, D. Taralla, D. Ernst, and R. Fonteneau, "Deep reinforcement learning solutions for energy microgrids management," in Proc. Eur. Workshop Reinf. Learn. (EWRL), Barcelona, Spain, 2016.
29. S. Thrun and T. M. Mitchell, "Lifelong robot learning," Robot. Auton. Syst., vol. 15, no. 1–2, pp. 25–46, 1995.
30. M. McCloskey and N. J. Cohen, "Catastrophic interference in connectionist networks: The sequential learning problem," Psychol. Learn. Motiv., vol. 24, pp. 109–165, 1989.
31. A. A. Rusu et al., "Progressive neural networks," arXiv preprint arXiv:1606.04671, 2016