



**HAL**  
open science

# Reinforcement learning for cooling rate control during quenching

Elie Hachem, Abhijeet Vishwasrao, Maxime Renault, Jonathan Viquerat,  
Philippe Méliga

► **To cite this version:**

Elie Hachem, Abhijeet Vishwasrao, Maxime Renault, Jonathan Viquerat, Philippe Méliga. Reinforcement learning for cooling rate control during quenching. *International Journal of Numerical Methods for Heat and Fluid Flow*, 2024, 34 (8), pp.3223-3252. 10.1108/HFF-11-2023-0713 . hal-04750779

**HAL Id: hal-04750779**

**<https://hal.science/hal-04750779v1>**

Submitted on 23 Oct 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Reinforcement learning for cooling rate control during quenching

---

## Abstract

In the process of quenching heat treatment, it is critical to establish the optimal process parameters producing the least residual stress magnitudes and related distortions and/or cracking, to reduce the cost of manufacturing high-quality components with intricate and durable designs and meet the stringent requirements of a broad range of high-performance industries. Because such effects occur as the result of uneven cooling in different regions of the quenched part, a feasible control objective is thus to enhance the spatial uniformity of heat removal, to prevent spatial gradients of irreversible strains typically originating from heterogeneous plastic deformation. For decades this process has been largely driven by trial and error, intuition and experience. In this study, a single-step Deep Reinforcement Learning (DRL) algorithm is used to provide the best possible cooling rate in industrial quenching processes governed by coupled pseudo-compressible Navier–Stokes and heat equations, along with latent heat formulation. The numerical reward fed to the neural network is computed with an in-house stabilized finite elements environment combining variational multi-scale (VMS) modeling of the governing equations, immerse volume method, and multi-component anisotropic mesh adaptation. A case of Rayleigh–Bénard convection in a high-aspect ratio, closed cavity is used first as testbed for the proposed methodology. In a second phase, we tackle several quenching numerical experiments aiming at improving temperature homogeneity within two-dimensional components in various shapes, whose results showcase the potential of DRL to produce unanticipated solutions by learning the effect of highly unsteady boiling flow physics on the temperature distribution.

*Keywords:* Deep Reinforcement Learning; Quenching; Computational Fluid Dynamics; Thermal Control, Multiphase Boiling

---

## 1. Introduction

Quenching is a heat treatment process used to modify the mechanical properties of steel materials. It consists in heating a part to achieve specific microstructure and material properties (*e.g.*, hardness, strength), after which the part is quickly cooled through the austenitizing temperature of a liquid quenchant that can be water, oil or brine, depending on the material and the desired properties. High cooling rates in the quenching process suppress the diffusion-controlled phase transformations and promote non-diffusional phase transformations to form martensite, one of the desired phases in quenched steel [1]. Although probably the oldest heat treatment process used by man to harden and strengthen steel, quenching remains vital as a safety procedure, to improve product durability and performance in various heavy industries with tight tolerances and high process repeatability requirements. Typical examples include the energy sector (*e.g.*, to manufacture seamless rolled rings), the automotive and aerospace industries (rings, gears, shafts and other transmission parts) or the construction industry (to avoid distortions in rods and bars).

Quenching effectiveness hinges in large measure on the ability of the quenchant to achieve maximal cooling without heating up. Meanwhile, in the absence of a systematic approach to select optimal process parameters, steel materials with varying compositions can respond differently to the hardening treatment, resulting in defects, rejections, reworks and added costs [2]. It is widely accepted that thermal expansions caused by large temperature gradients (which the present study focuses on), volumetric expansions caused by martensitic transformations, or the combination of both produce high residual stresses, contributing to distortions, cracking and fractures that all negatively impact the microstructure of the part and lead to reduced fatigue strength [3]. Despite a large body of literature on this topic contributing to advancements in manufacturing processes, failures persist in the quenching process, with heavy or intricately shaped components having

24 the highest rejection percentages [4]. In the present context of tackling climate change while  
25 ensuring sustainable growth, the capability of optimally controlling the outcome of quenching is  
26 more relevant than ever, to deliver durable, high-quality parts at manageable costs while minimizing  
27 the many types of waste that can occur during manufacturing [5].

28 Quenching involves heating the component at temperatures of the order of 1000°C. The initial  
29 temperature across any given geometry is constant [1], and the temperature distribution during  
30 cooling depends solely upon the outward heat flux from the component, and thereby upon highly  
31 localized heat transfer coefficients. The complexity of the quenching process is related to the boiling  
32 heat transfer and phase transformations occurring on the surface of the immersed part (mostly  
33 the nucleation phase), that makes it very sensitive to small variations in process parameters.  
34 For instance, for a part to have uniform heat transfer, any vapor pockets must be minimized  
35 for the quenching medium to wrap perfectly around its surface. This requires identifying relevant  
36 process parameters, a real challenge that requires a perfect balance in the choice of (among many  
37 other factors) the austenitizing temperature, the immersion rate and the orientation of the treated  
38 component, and the temperature, volume and agitation of the quenching medium. Actually, this  
39 must be repeated for each particular geometry, as even the slightest asymmetry in the shapes,  
40 combined with tiny variations in process parameters, can add to the lack of uniformity and cause  
41 the appearance of unwanted residual stresses associated with diminished mechanical properties [6].

42 For decades, process parameters have been essentially adjusted through intuition, trial and  
43 error, and professional experience. The premise of this research is that this highly complicated  
44 task can be rationally and efficiently solved using reinforcement learning (RL), a machine learning  
45 method for solving sequential decision-making problems. While RL has long been limited to low  
46 dimensional problems, several major obstacles have been lifted using the feature extraction capabilities  
47 of deep neural networks and their ability to handle high-dimensional state spaces, giving rise to  
48 Deep Reinforcement Learning (Deep RL or DRL). This has yielded unprecedented efficiency after  
49 short training spans in many domains such as robotics [7], language processing [8] or games [9, 10],  
50 but DRL is also used in many industrial applications, including autonomous cars [11], or data  
51 center cooling [12]. The potential application in fluid mechanics is also highly promising, for which  
52 efforts are ongoing thanks to the sustained commitment from the machine learning community.  
53 A few dozen studies provided insight into the performance improvements to be delivered, with  
54 particular focus on shape optimization and flow control applied to drag reduction; see [13, 14] for  
55 recent reviews. Meanwhile, the literature on thermal control is scarce, with (at time of writing)  
56 only a handful of studies applying DRL-based approaches to control natural [15-17] and forced  
57 convection [16, 18, 19], although DRL has become a quickly emerging topic in a wide range of  
58 thermal applications, from the shape optimization of heat exchangers [20] to the implementation  
59 of thermal digital twins [21], including energy efficiency in civil engineering [22] and the estimation  
60 of effective statistical properties in complex media [23].

61 This work aims at introducing DRL into the field of quenching control. It stands as a follow-up  
62 of our previous work on DRL-based control of forced convection [16, 18], and combines DRL with  
63 advanced immersed methods for the simulation of liquid-vapor phase change, to extend the scope to  
64 boiling and increase the complexity of the targeted applications. The proposed framework leverages  
65 the capacity of neural networks to accurately approximate the mapping function between input  
66 and output spaces, as well as the dynamic programming inherent in the reinforcement learning  
67 algorithm. There is no similar study in the literature, to the best of our knowledge, for which  
68 a possible explanation is the lack of tried and tested computational solvers capable of reliably  
69 simulating the quenching of solid parts. In this context, to act as DRL agent, we use the single-  
70 step Proximal Policy Optimization (single-step PPO) algorithm introduced in [24, 25], intended  
71 for optimization and open-loop control problems whose optimal policy to be learnt by the agent  
72 is state-independent (in which case it suffices to update the neural network parameters only once  
73 per episode). The reward function used to train our PPO agent is computed by an in-house, high-  
74 fidelity computational framework that predicts accurately the boiling heat transfer behavior of a  
75 liquid in the near field of a heated immersed solid, while taking into account the gas-liquid phase  
76 changes, the vapor formation and their dynamics, and ultimately, the cooling of the solid [26, 27].  
77 The effectiveness of the DRL-CFD approach is demonstrated by evaluating the homogeneity of the  
78 cooling effect on various systems comprising a 2-D part with controlled orientation in a liquid tank.  
79 Of note, this is a proof of concept study to lay the groundwork for future research in the field,

80 not a comparable study demonstrating the competitiveness of DRL-CFD by benchmarking the  
 81 performance against generic models available off-the-shelf. We therefore focus on positioning the  
 82 method as an efficient and theoretically well-founded tool to find the best process parameters and  
 83 achieve the desired final characteristics, and exchange views on the main challenges that should be  
 84 considered to realize its potential for application in real-world manufacturing environments.

85 With these considerations in mind, the paper is structured as follows: in section 2, we outline  
 86 the main features of the finite element CFD environment used to simulate the quenching process  
 87 and compute the numerical reward fed to the DRL agent while taking into account liquid-vapor-  
 88 solid interactions with boiling heat transfer and liquid-vapor transformations. Section 3 breaks  
 89 down the baseline principles of DRL and PPO, together with the specifics of the single-step PPO  
 90 algorithm. Section 4 provides the particulars of DRL experiments and implementation, and revisits  
 91 a Rayleigh–Bénard convection case adapted from [16] for the purpose of validation and assessment  
 92 part of the method capabilities. In, section 5, DRL is used to improve the temperature homogeneity  
 93 within various parts immersed in a liquid tank, whose geometrical complexity ranges from the  
 94 simplest rectangular brick shape to intricate, irregular shapes of engineering interest. Conclusion  
 95 and strategies for future works in this area are presented in section 6.

## 96 2. Computational fluid dynamics and phase change solver

97 This section breaks down the main ingredients of the adaptive Eulerian framework for the  
 98 simulation of both boiling and evaporation phenomena occurring at the interface of a heated 2-D  
 99 solid immersed in a liquid tank. Exhaustive derivation and implementation details are provided in  
 100 Ref. [27], to which the interested reader is referred for further deepening. Suffice it to say here that  
 101 a pseudo-compressible model accounting for mass transfer at the liquid-vapor interface is solved  
 102 with the Immerse Volume Method [28, 29], which allows to compute the heat transfer between the  
 103 solid and the (liquid) quenchant from the individual material properties on either side of it. This  
 104 in turn obviates the need to compute a heat transfer coefficient, which would have been a limiting  
 105 issue from the present numerical experiments where varying the shape and orientation of the solid  
 106 is integral to the optimization process.

### 107 2.1. Interface capturing method

108 The level set method is employed as an interface capturing technique to identify and monitor  
 109 the evolution of the liquid-phase interface using the zero iso-value of a smooth level set function.  
 110 This function distributes the corresponding physical properties in space according to a mixing  
 111 law. Let  $\Omega$  represent the entire domain, and  $\Omega_l$  and  $\Omega_v$  denote the liquid and vapor domains,  
 112 respectively. The level set function is a signed distance function from the interface  $\Gamma = \Omega_l \cap \Omega_v$ ,  
 113 defined at each node  $X$  as follows:

$$\phi = \begin{cases} -\text{dist}(X, \Gamma) & \text{if } X \in \Omega_l, \\ 0 & \text{if } X \in \Gamma, \\ \text{dist}(X, \Gamma) & \text{if } X \in \Omega_v, \end{cases} \quad (1)$$

114 with the convention that  $\phi > 0$  in the vapor domain. In the absence of mass transfer between the  
 115 liquid and vapor phases, the evolution of the level set is governed by the transport equation.

$$\partial_t \phi + \mathbf{u} \cdot \nabla \phi = 0, \quad (2)$$

116 with  $\mathbf{u}$  a velocity. The level set, defined as a distance function, satisfies  $\|\nabla \phi\| = 1$ . However, it can  
 117 lose this property during the convection process, in which case it requires reinitialization to pre-  
 118 vent numerical instabilities. A common method for reinitialization is solving the Hamilton–Jacobi  
 119 equation

$$\partial_\tau \phi + s(\phi)(\|\nabla \phi\| - 1) = 0, \quad (3)$$

120 where  $\tau$  is a pseudo time-step and  $s(\phi)$  is the sign function of  $\phi$ .

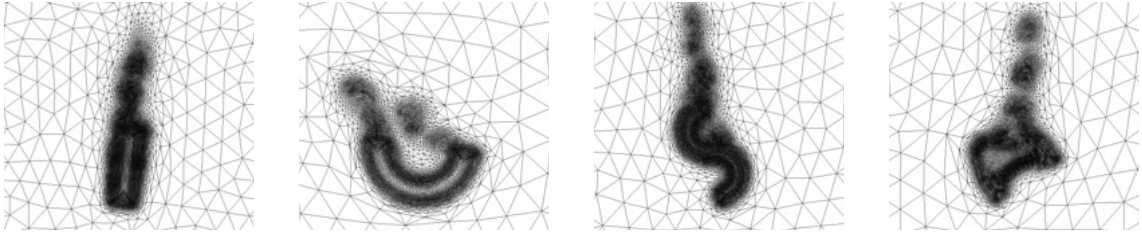


Figure 1: Examples of anisotropic meshes used to simulate the boiling phenomenon.

121 Once the level set function is computed, the individual physical properties (for example  $\rho_l$  and  
 122  $\rho_v$ , respectively the liquid and vapor densities) are distributed using a mixing law according to

$$\rho = \rho_l(1 - H_\epsilon(\phi)) + \rho_v H_\epsilon(\phi), \quad (4)$$

123 where  $H_\epsilon$  is the smoothed Heaviside function defined as

$$H_\epsilon(\phi) = \begin{cases} 0 & \text{if } \phi < -\epsilon, \\ \frac{1}{2}\left(1 + \frac{\phi}{\epsilon} + \frac{1}{\pi} \sin\left(\pi \frac{\phi}{\epsilon}\right)\right) & \text{if } |\phi| \leq \epsilon, \\ 1 & \text{if } \phi > \epsilon, \end{cases} \quad (5)$$

124 and  $\epsilon = 2h_\infty$  is a regularization parameter set here to twice the mesh size in the normal direction  
 125 to the interface.

## 126 2.2. Anisotropic mesh adaptation

127 Difficulties may arise in an immersed multiphase framework due to the discontinuities in mate-  
 128 rial properties between the solid and fluid regions. These discontinuities are particularly challenging  
 129 when they intersect the mesh elements arbitrarily, potentially compromising the accuracy of the  
 130 solution or causing it to fail entirely. To address this, we employ the anisotropic mesh adaptation  
 131 technique described in [30]. This approach generates highly stretched, well-oriented elements and  
 132 distributes the fluid properties as accurately and smoothly as possible over a minimal thickness  
 133 around the interface. This allows for the effective capture of sharp gradients at a low computational  
 134 cost, ensuring consistency between the solution anisotropy and the mesh. This is achieved by cal-  
 135 culating modified distances from a metric, a symmetric positive-definite tensor whose eigenvectors  
 136 define preferential directions along which mesh sizes are determined based on the corresponding  
 137 eigenvalues. This metric is isotropic far from the interface, with the mesh size set to  $h_\infty$  in all  
 138 directions. However, near the interface, the metric becomes anisotropic, with the mesh size set to  
 139  $h_\perp$  in the direction normal to the liquid/vapor interface and to  $h_\infty$  in the other directions. For a  
 140 desired thickness  $\delta$ , this can be expressed as follows:

$$\mathbf{M} = K(\phi)\mathbf{n} \otimes \mathbf{n} + \frac{1}{h_\infty^2}\mathbf{I} \quad \text{with} \quad K(\phi) = \begin{cases} 0 & \text{if } |\phi| \geq \delta/2, \\ \frac{1}{h_\perp^2} - \frac{1}{h_\infty^2} & \text{if } |\phi| < \delta/2, \end{cases} \quad (6)$$

141 where  $\mathbf{n} = \nabla\phi/|\nabla\phi|$  is the normal to the interface deduced from the gradient of the level set.  
 142 An a posteriori anisotropic error estimator is then used to minimize the interpolation error while  
 143 maintaining a fixed number of edges in the mesh. This is done using multi-component error vectors  
 144 taking into account the gradients of multiple scalar and/or vector fields [30–33]. In the following  
 145 numerical experiments, adaptivity combines velocity components and magnitude, temperature and  
 146 level set, all normalized by their respective global maximum to ensure that a field much larger in  
 147 magnitude does not dominate the error estimator. Examples of the adapted meshes generated  
 148 in this study are shown in figure 1. These examples illustrate the appropriate refinement and  
 149 deformation of mesh elements, which are extremely fine and elongated near the interfaces between  
 150 the solid, liquid, and vapor, but coarse and uniform away from the interfaces.

151 *2.3. Phase change model*

152 The multiphase framework used to simulate phase change problems relies on pseudo-compressible  
153 mass and momentum conservation written in the form of modified Navier–Stokes equations

$$\nabla \cdot \mathbf{u} = \dot{m} \left( \frac{1}{\rho_v} - \frac{1}{\rho_l} \right) |\nabla \phi| \delta_\epsilon(\phi), \quad (7)$$

$$\rho(\partial_t \mathbf{u} + \mathbf{u} \cdot \nabla \mathbf{u}) - \nabla \cdot (2\mu \boldsymbol{\varepsilon}(\mathbf{u})) + \nabla p = \mathbf{f}_\gamma + \rho \mathbf{g}. \quad (8)$$

154 Here,  $\rho$  represents the density,  $\mu$  the dynamic viscosity,  $\mathbf{u}$  the velocity,  $p$  the pressure,  $\boldsymbol{\varepsilon}(\mathbf{u})$  the rate  
155 of deformation tensor,  $\mathbf{g}$  the gravity acceleration, and we note that the velocity is not divergence-  
156 free, since the continuity equation is forced by a surface mass transfer rate, denoted as  $\dot{m}$ , that  
157 measures the exchange of mass occurring at the liquid/vapor interface. Additionally,  $\delta_\epsilon$  is the Dirac  
158 function

$$\delta_\epsilon(\phi) = \begin{cases} \frac{1}{2\epsilon} \left( 1 + \cos \left( \pi \frac{\phi}{\epsilon} \right) \right) & \text{if } |\phi| \leq \epsilon, \\ 0 & \text{if } |\phi| > \epsilon, \end{cases} \quad (9)$$

159 locating the interface and smoothed with the same regularization parameter  $\epsilon$  as the Heaviside  
160 function (5). Lastly,  $\mathbf{f}_\gamma$  is a body force representing the impact of surface tension, formulated  
161 within the framework of the Continuum Surface Force [34] as

$$\mathbf{f}_\gamma = -\gamma \kappa \delta_\epsilon(\phi) \mathbf{n}, \quad (10)$$

162 where  $\gamma$  is the surface tension coefficient and  $\kappa = \nabla \cdot \mathbf{n}$  is the mean interface curvature. Likewise,  
163 energy conservation is expressed in the form of a modified heat equation

$$\rho c_p (\partial_t T + \mathbf{u} \cdot \nabla T) = \nabla \cdot (\lambda \nabla T) - (\mathcal{L} + (c_{p_v} - c_{p_l})(T - T_{sat})) \dot{m} \delta_\epsilon(\phi) |\nabla \phi| \frac{\rho^2}{\rho_l \rho_v}. \quad (11)$$

164 Here,  $T$  denotes the temperature,  $T_{sat}$  the saturation temperature,  $\mathcal{L}$  the liquid latent heat of  
165 vaporization,  $c_p$  the specific heat,  $c_{p_l}$  (resp.  $c_{p_v}$ ) the specific heat in the liquid (and vapor) phases  
166 respectively, and  $k$  the thermal conductivity. Due to mass transfer, the level set equation (1) is  
167 consequently modified into

$$\partial_t \phi + \left( \mathbf{u} - \frac{\rho}{\rho_l \rho_v} \dot{m} \frac{\nabla \phi}{|\nabla \phi|} \right) \cdot \nabla \phi = 0, \quad (12)$$

168 for the interface to be convected not only by the Navier–Stokes velocity, but also the velocity of  
169 the vapor front. The mass transfer rate  $\dot{m}$  in the aforementioned general formulation is computed  
170 based on the balance of fluxes at the interface, hence

$$\dot{m} = \frac{\int_{\Omega_i} \delta_\epsilon(\phi) (-k_v \nabla T_v + k_l \nabla T_l) \cdot \mathbf{n} d\Omega_i}{\int_{\Omega_i} \delta_\epsilon(\phi) d\Omega_i}, \quad (13)$$

171 where we sum over all elementary volumes  $\Omega_i$  intersected by the interface.

172 *2.4. Variational multiscale modeling*

173 All equations are solved using equal-order linear approximations for the velocity and pressure  
174 variables, for which we employ stabilized weak forms cast in the Variational Multiscale (VMS)  
175 framework. This approach enhances the stability of the Galerkin method by introducing addi-  
176 tional integrals over the element interior. It effectively mitigates the node-to-node oscillations  
177 that typically arise when discretization schemes violate the Babuska–Brezzi condition. The fun-  
178 damental concept involves decomposing all quantities into large and small-scale components, rep-  
179 resenting different levels of resolution. The effect of the unresolved small-scale details, beyond the  
180 finite element mesh resolution, is approximated on the large scale through consistently derived  
181 residual-based terms. Extensive validation and verification of this numerical framework accuracy  
182 and reliability are detailed in [27]. Interested readers are encouraged to consult this reference for  
183 comprehensive information on the VMS formulations, stabilization parameters, and discretization  
184 schemes applied to the phase-change model.



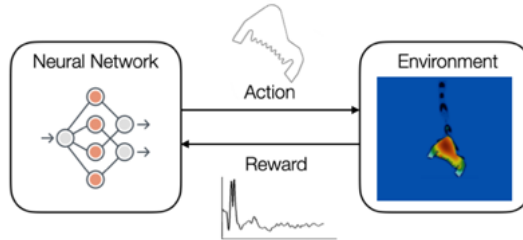


Figure 2: Sketch of the present single-step PPO action-loop. The quenching CFD environment with phase change and the DRL agent are coupled two-way through actions and rewards. At each episode, the same input state  $s_0$  is provided to the agent, which in turn provides  $n$  actions to  $n$  parallel environments. The latter return  $n$  rewards, that evaluate the quality of each action taken. Once all the rewards are collected, an update of the agent parameters is made using the PPO loss [14].

### 185 3. Deep reinforcement learning and single-step proximal policy optimization

186 In the following, quenching processes are optimized by solving a decision-making problem with  
 187 reinforcement learning (RL), a method by which an agent learns to maximize rewards through  
 188 trial-and-error interactions with its environment. At each step, the agent observes the state  $s_t$  of  
 189 the environment and takes an action  $a_t$ , which leads to a reward  $r_t$  and a transition to the next  
 190 state  $s_{t+1}$ . This repeats until a termination state is reached, the primary goal of the agent being  
 191 to learn a sequence of actions that maximizes its cumulative reward over an episode (which is the  
 192 reference unit for agent update). In the present context, the environment is a CFD simulation with  
 193 phase change, that uses the stabilized finite element framework described above. The agent is a  
 194 RL-trained neural network coupled two-way with the environment, as illustrated in Fig. 2 on the  
 195 one hand, the actions sampled by the agent are used to generate the workpieces meshes immersed  
 196 in the CFD simulation. On the other hand, the reward function needed by the agent to learn (here,  
 197 a measure of cooling homogeneity) is obtained by post-processing of the CFD data.

#### 198 3.1. Single-step deep reinforcement learning

199 In deep reinforcement learning (DRL), the agent is implemented as a neural network, most  
 200 often structured into a series of fully connected layers, with information flowing from the input  
 201 layer to the output layer through hidden layers, each of which acts as a function from  $\mathbb{R}^m$  to  
 202  $\mathbb{R}^n$ . The neural network learns the relationship between input (action) and output (reward) data  
 203 by iteratively adjusting the weights and biases through a process known as back-propagation,  
 204 which moves from the output layer back through the hidden layers and to the input layer. This  
 205 training process enables the network to refine its predictions and improve performance over time.  
 206 In classical DRL, network updates are performed after multi-step episodes for the agent to learn the  
 207 set of actions  $a^*$  yielding the highest possible reward. The present approach is conversely cast in the  
 208 single-step deep reinforcement learning framework, an approach that has emerged from the premise  
 209 that network updates can be performed after one-step episodes (of single-step episodes, hence by  
 210 extension, single-step DRL) if the optimal behavior to be learned is independent of state. A single-  
 211 step DRL agent learns instead the optimal mapping  $f_{\theta^*}$  such that  $a^* = f_{\theta^*}(s_0)$ , where  $s_0$  is an  
 212 input state, usually a constant vector, repeatedly passed to the agent (hence the stateless moniker).  
 213 A significant advantage of single-step DRL is that it allows to use much smaller networks compared  
 214 to the typical architectures used in traditional DRL approaches. This is because the agent does  
 215 not need to learn a complex state-action relationship but only the transformation from a constant  
 216 input state to a specific action.

#### 217 3.2. Single-step proximal policy optimization

218 In the following, a neural network is trained with single-step Proximal Policy Optimization, the  
 219 single-step variant of the ubiquitous PPO RL algorithm introduced in Ref. [24, 25] and shown in  
 220 Ref. [16] to hold potential as a reliable, go-to black-box optimizer for natural and forced convection

221 heat transfer enhancement. In short, one neural network outputs the mean and variance of a  $d$ -  
 222 dimensional multivariate normal distribution (with  $d$  the dimension of the action required by the  
 223 environment). All variables are assumed to have equal variance and to be uncorrelated, meaning  
 224 that the covariance matrix is identity, thereby establishing an isotropic sampling region for the  
 225 upcoming episode. Actions drawn in  $[-1, 1]^d$  are then mapped into relevant physical ranges, a step  
 226 deferred to the environment as being problem-specific. Just like PPO, the algorithm computes  
 227 adaptive learning rates for each policy parameter based on the gradient of the loss function

$$\mathbb{E}_{a \sim \pi_\theta} \left[ \min \left( \frac{\pi_\theta(a)}{\pi_{\theta_{old}}(a)}, 1 + \epsilon \operatorname{sgn}(\widehat{A}^{\pi_\theta}(a)) \right) \widehat{A}^{\pi_\theta}(a) \right], \quad (14)$$

228 where  $\widehat{A}^{\pi_\theta}$  serves as a biased estimator of the advantage function  $A^{\pi_\theta}$ , quantifying the convenience  
 229 of taking action  $a$  compared to the average value (normalized to zero mean and unit variance).  
 230 The  $\epsilon$  parameter defines a clipping range that limits how much the new policy can deviate from the  
 231 old policy. A negative (positive) advantage decreases (increases) the likelihood of taking action  $a$ ,  
 232 but always within a proportion smaller than  $\epsilon$ . If this threshold is exceeded, the policy change is  
 233 constrained by a ceiling of  $1 + \epsilon$  (or a floor of  $1 - \epsilon$ ), enforced by the minimization operation in Eq. (1)  
 234 and its corresponding argument. This cautious approach inherited from the parent algorithm [35]  
 235 ensures that the current and updated policies exhibit similar behavior, which prevents the agent  
 236 from making abrupt policy changes that could lead to significant performance deterioration. A  
 237 nuanced distinction between PPO and its single-step variant that is worth mentioning is that PPO  
 238 operates as an actor-critic framework, where an actor network learns the policy and a critic network  
 239 estimates advantages. In contrast, single-step PPO does not rely on critic evaluations (thus not  
 240 following the actor-critic paradigm), as it involves only a single state-action pair trajectory. Setting  
 241 the discount factor  $\gamma = 1$  adjusts the balance between immediate and future rewards, simplifying  
 242 the advantage to the whitened reward, as further explained in Ref. [25].

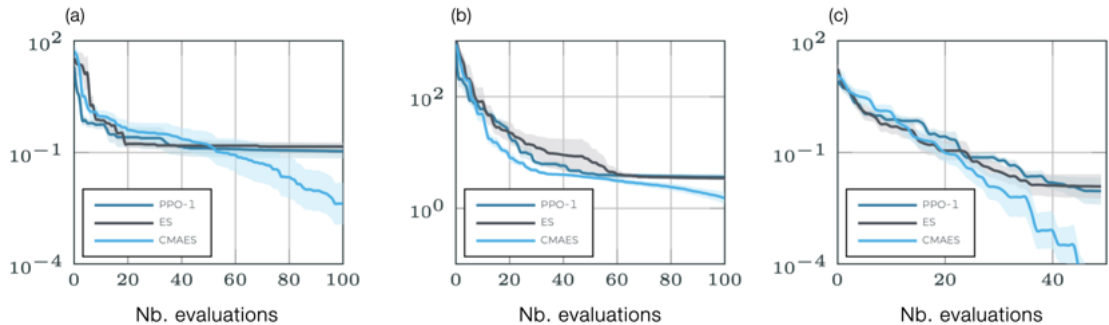


Figure 3: **Benchmark minimization problems** for the (a) two- and (b) five-dimensional Rosenbrock functions, and (c) the two-dimensional Branin function, using the present single-step PPO algorithm and reference  $(\mu-\lambda)$ -ES and CMA-ES evolutionary algorithms.

243 For context, the convergence properties are illustrated in figure [3] for minimization test cases  
 244 of two- and five-dimensional Rosenbrock functions, whose global minimum is notoriously difficult  
 245 to catch, and two-dimensional Branin function, that has two identical global minima. Single-  
 246 step PPO is benchmarked against classical  $(\mu - \lambda)$ -ES and CMA-ES evolutionary methods, all  
 247 implemented in in-house production codes. The initial parameters and starting points are identical  
 248 for all methods to ensure a fair comparison. All runs are afforded the same budget, namely 500  
 249 evaluations for Rosenbrock (20 episodes with 5 parallel environments per episode in PPO vs. 20  
 250 generations with 5 individuals per generation in ES algorithms) and 50 evaluations for Branin (10  
 251 episodes with 5 parallel environments per episode in PPO vs. 10 generations with 5 individuals per  
 252 generation in ES). A large initial standard deviation is used by default, to ensure a good exploration  
 253 of the optimization domain. Finally, in order to emphasize flexibility and generalizability, all PPO  
 254 runs are tackled without fine-tuning of the algorithm, so all runs use the same meta-parameters,  
 255 namely two steps mini-batches to update the network for 32 epochs, with learning rate set to



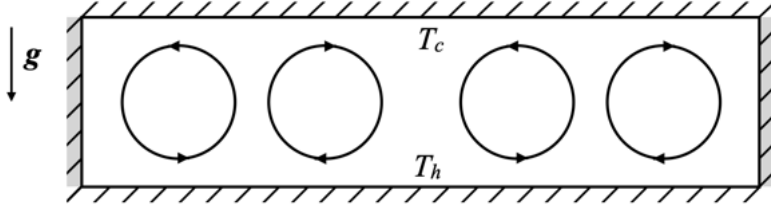


Figure 4: Schematic of the 2-D differentially heated cavity set-up. The gray shade indicates the insulated walls.

256 0.005 and PPO loss clipping range to  $\epsilon = 0.2$ . Performances are averaged over 10 runs, with  
 257 standard deviations shown as the light shade around. Unsurprisingly, CMA-ES performs best,  
 258 which reflects the improvement of efficiently elongating the research area to suit the shape of the  
 259 cost function. Among isotropic exploration methods, single-step PPO achieves final cost levels  
 260 similar to  $(\mu - \lambda)$ -ES, with faster convergence and better performance at intermediate stages (the  
 261 final performance level ultimately saturates for the Rosenbrock function because the minimum is  
 262 in a long, narrow valley, and PPO and  $(\mu - \lambda)$ -ES use isotropically sampled approximations of the  
 263 descent direction. The general conclusion is that 1. single-step PPO exhibits strong performance  
 264 compared to methods relying on similar isotropic search distributions, and 2. it is imperative to  
 265 utilise anisotropic search distributions to outperform more advanced methods on a consistent basis,  
 266 an issue that is being addressed in current research efforts by the authors [36].

## 267 4. Control of Rayleigh–Bénard convection

### 268 4.1. Case description

269 In order to assess the method’s capability compared to data available in the literature, this  
 270 section tackles the control of Rayleigh–Bénard (natural) convection in the two-dimensional dif-  
 271 ferentially heated cavity sketched in figure 4(a). This is a widely studied benchmark system for  
 272 thermally-driven flows, relevant in nature and technical applications (*e.g.*, ocean and atmospheric  
 273 convection, materials processing, metallurgy), that is thus suitable to assess relevance of the nu-  
 274 merical framework. The canonical initial condition is a fluid at rest that is being heated from the  
 275 lower wall and/or cooled from the upper wall, with natural convection ensuing as a result of the  
 276 induced temperature gradients and fluid-buoyancy effects. A Cartesian coordinate system is used  
 277 with origin at the lower-left edge, horizontal  $x$ -axis, and vertical  $y$ -axis. The vertical sidewalls are  
 278 perfectly insulated from the outside (adiabatic). The horizontal walls are isothermal: the upper  
 279 wall is kept at a constant “cold” temperature  $T_c$ , and the lower wall is entirely controllable via a  
 280 space-varying “hot” distribution  $T_h(x)$  such that  $\langle T_h \rangle > T_c$ , where the brackets denote the average  
 281 over space the  $x$ -position along the hot wall. Several studies have reported the benefits of similarly  
 282 using DRL for natural convection heat transfer performance in laterally and bottom-heated square  
 283 cavities [15, 16]. Here, a laterally extended domain with aspect ratio 4:1 is considered, for which  
 284 the rationale is twofold: first, the convection cells are closer to the unconstrained cells obtained  
 285 in wide domains relevant for industrial use. Second, control is more demanding, as it makes more  
 286 difficult for the DRL agent to use the walls to move around and break the cells.

287 In this case of no phase change, the governing equations are the classical Navier–Stokes and  
 288 heat equations written under the Boussinesq approximation as

$$\nabla \cdot \mathbf{u} = 0, \quad (15)$$

$$\rho(\partial_t \mathbf{u} + \mathbf{u} \cdot \nabla \mathbf{u}) - \nabla \cdot (2\mu \boldsymbol{\varepsilon}(\mathbf{u})) + \nabla p = -\rho\beta(T - T_c)\mathbf{g}, \quad (16)$$

$$\rho c_p(\partial_t T + \mathbf{u} \cdot \nabla T) = \nabla \cdot (\lambda \nabla T), \quad (17)$$

289 where  $\beta$  is the thermal expansion coefficient, and we use  $T_c$  as Boussinesq reference temperature.  
 290 The above equations are solved assuming no-slip on the walls and temperature boundary conditions

$$\partial_x T(0, y, t) = \partial_x T(4H, y, t) = 0, \quad T(x, 0, t) = \langle T_h \rangle + \tilde{T}_h(x), \quad T(x, H, t) = T_c, \quad (18)$$

291 where  $\tilde{T}_h$  is a zero-mean (in the sense of the space-average) hot temperature fluctuation, whose  
 292 magnitude is bounded according to

$$|\tilde{T}_h(x)| \leq \Delta T_{max}, \quad (19)$$

293 to avoid extreme and nonphysical temperature gradients. This system is controlled by the Rayleigh  
 294 and Prandtl numbers, set here to  $Ra = 10^4$  and  $Pr = 0.71$  (air value at room temperature) using  
 295 the cavity height, the heat conductivity time, and the (time and space-constant) difference between  
 296 the mean horizontal temperatures as reference scales.

#### 297 4.2. Control and reward

298 Following [15, 16], we seek to optimize the distribution of hot temperature fluctuations  $\tilde{T}_h$   
 299 by training a DRL agent in selecting piece-wise constant temperatures over  $4n_s$  identical seg-  
 300 ments (labeled from left to right), each of which allows only two pre-determined states referred  
 301 to as hot or cold. This is intended to reduce the complexity and the computational resources, as  
 302 large/continuous action spaces are known to be challenging for the convergence of RL methods [37].  
 303 All results reported herein are for  $\Delta T_{max} = 0.75$ , for the hot temperature to vary in the range  
 304 from 0.25 to 1.75.

305 Based on the topology of the baseline, uncontrolled solution (more details in the following),  
 306 symmetric actuation with respect to the vertical centerline is used, in which the network outputs  
 307  $n_s$  discrete values  $\hat{T}_{h,k \in \{1 \dots n_s\}} = \pm \Delta T_{max}$ , mapped into actual physical fluctuations over the first  
 308  $n_s$  segments using

$$\tilde{T}_{h,k} = \frac{\hat{T}_{h,k} - \langle \hat{T}_{h,k} \rangle}{\max_{l \in \{1 \dots n_s\}} \left( 1, \frac{|\hat{T}_{h,l} - \langle \hat{T}_{h,l} \rangle|}{\Delta T_{max}} \right)}, \quad (20)$$

309 to fulfill the zero-mean and upper bound constraints. This pattern is mirrored over the  $n_s$  following  
 310 segments, then repeated over the next  $2n_s$  segment, according to

$$\tilde{T}_{h,k} = \begin{cases} \tilde{T}_{h,2n_s+1-k}, & \text{if } n_s < k \leq 2n_s, \\ \tilde{T}_{h,k-2n_s}, & \text{if } 2n_s < k \leq 3n_s, \\ \tilde{T}_{h,4n_s-k+1}, & \text{if } 3n_s < k \leq 4n_s. \end{cases} \quad (21)$$

311 We avoid nonphysical sharp discontinuities across segments using hyperbolic tangent functions to  
 312 regularize the temperature fluctuation on each segment relatively to its immediate neighbors. This  
 313 ensures continuously differentiable actuation from one segment to another, and is accounted for  
 314 in practice to impose the zero-mean condition from (20)-(21). We set here  $n_s = 10$ , a spatial  
 315 granularity of 10 segments per vortex of the uncontrolled solution, that has been found to allow  
 316 suitable controllability.

317 The agent is incentivized to alleviate convective heat transport by receiving the reward  $r_t =$   
 318  $-\text{Nu}$ , where  $\text{Nu}$  is the Nusselt number defined as the non-dimensional temperature gradient av-  
 319 eraged over the hot bottom wall (hence  $r_t = 1$  if heat transfer is by pure conduction, and  $r_t > 1$   
 320 otherwise). All values are computed from 400 points (100 per vortex of the uncontrolled solution)  
 321 uniformly distributed along the hot bottom wall. A typical DRL simulation runs on 64 CPU cores,  
 322 using 8 environments of 8 CPU each. The agent is a fully connected network with two hidden layers  
 323 holding two neurons. The resolution process uses eight environments and two steps mini-batches  
 324 to update the network for 32 epochs, with the learning rate set to 0.005 and PPO loss clipping  
 325 range to  $\epsilon = 0.2$ . The agent then generates an improved batch of actions for the next generation,  
 326 and the process repeats until convergence is achieved.

#### 327 4.3. Results

328 Using the numerical methods described in the previous sections, the uncontrolled solution is  
 329 computed starting from an initial condition of zero velocity and uniform temperature equal to  $T_c$ .  
 330 Every five time steps, anisotropic mesh adaptation is performed under the constraint of a fixed  
 331 number of elements  $n_{el} = 80000$ , using velocity and temperature as multiple-component adaptation

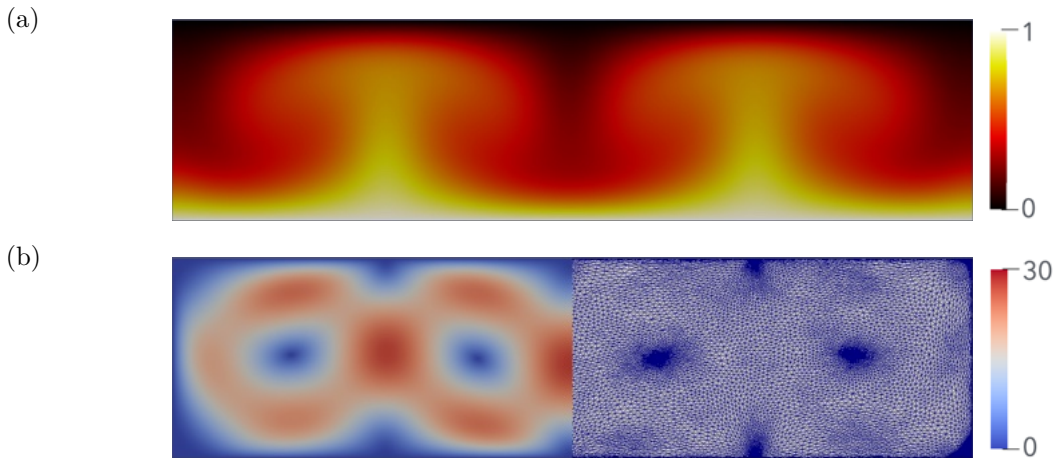


Figure 5: Uncontrolled twin-cell steady state solution. (a) Iso-contours of the temperature between 0 and 1. (b) Iso-contours of the velocity magnitude between 0 and 30, together with the corresponding anisotropic adapted mesh.

332 criterion (but no level-set, as the solid is solely at the boundary of the computational domain, where  
 333 either the temperature, or the heat flux is known). As shown in figure 5(a,b), the cold downwelling  
 334 and hot upwelling fluid organizes into a twin-cell configuration made up of two pairs of counter-  
 335 rotating vortices, that ultimately becomes stationary. This occurs after approximately 15 time  
 336 units, after which the Nusselt number converges to a value of  $Nu = 2.56$ . The corresponding  
 337 adapted mesh shown in the right half of figure 5(b) stresses that all boundary layers are sharply  
 338 captured via extremely stretched elements, and that the adaptation strategy yields refined meshes  
 339 near high temperature gradients and close to the side walls. Note however, the mesh refinement  
 340 is not only along the boundary layers but also close to the recirculation regions near the cavity  
 341 center, while the elements in-between are coarser and essentially isotropic.

342 Control runs comprise of 100 learning episodes, each of which marches in time the above baseline  
 343 initial state for a duration of 300 time units using time step  $\Delta t = 1$ . This represents 800 simulations  
 344 per run, each of which is performed on 8 cores and lasts 40mn, hence 530h of total CPU cost per  
 345 run (about 65h in wall time). It is out of the scope of this work to analyze in details the flow  
 346 patterns that develop when control is applied at the bottom of the cavity. Suffice it to say that the  
 347 outcome consistently exhibits twin-cell patterns of varying size and magnitude, accompanied by  
 348 corner eddies at the bottom of the cavity. This is best seen in figure 6 through several iso-contours of  
 349 the steady-state temperature, each of which corresponds to a different learning episode performed  
 350 over the course of the DRL optimization, and thus, to a different temperature distribution at  
 351 the hot bottom wall. For all cases, steady state is achieved within a few ten time units, but  
 352 the counter-rotating vortices must occasionally exchange place to suit the specifics of the control,  
 353 which may take up to a few hundred time units. Mesh adaptation is an asset in this regards, as  
 354 it allows to capture the anisotropy of the transient and asymptotic dynamics, intensified by the  
 355 sharp (albeit continuous) boundary conditions; see in figure 7 the detailed time-evolution of the  
 356 temperature field for the one control episode that yields the steady-state in figure 6(d), together  
 357 with the corresponding adapted meshes. We show in figure 8 the evolution of the reward (Nusselt  
 358 number), for which performances have been averaged over 5 independent runs, as is customary  
 359 in machine learning evaluation. The run-averaged mean Nusselt number during the optimization  
 360 process is shown as gray line, with standard deviations shown as the light shade around. Finally,  
 361 the black line shows the moving average Nusselt number, computed from the run-averaged mean  
 362 as the sliding average over the 50 latest reward values (or the whole sample if it has insufficient  
 363 size). The latter decreases monotonically and reaches a plateau after about 30 episodes, although  
 364 we notice that sub-optimal distributions keep being explored occasionally.

365 The optimal computed by averaging over the 10 latest episodes (hence the 800 latest instant  
 366 values) is 2.03, which corresponds to an efficiency of about 20% compared to the uncontrolled case.  
 367 As evidenced by the best temperature distribution over the 5 optimization runs in figure 9, this  
 368 requires providing a much wider plume by heating on either side of the vortex cores, although  
 369 convection ultimately remains, consistently with the results in Ref. [16] at the same Rayleigh

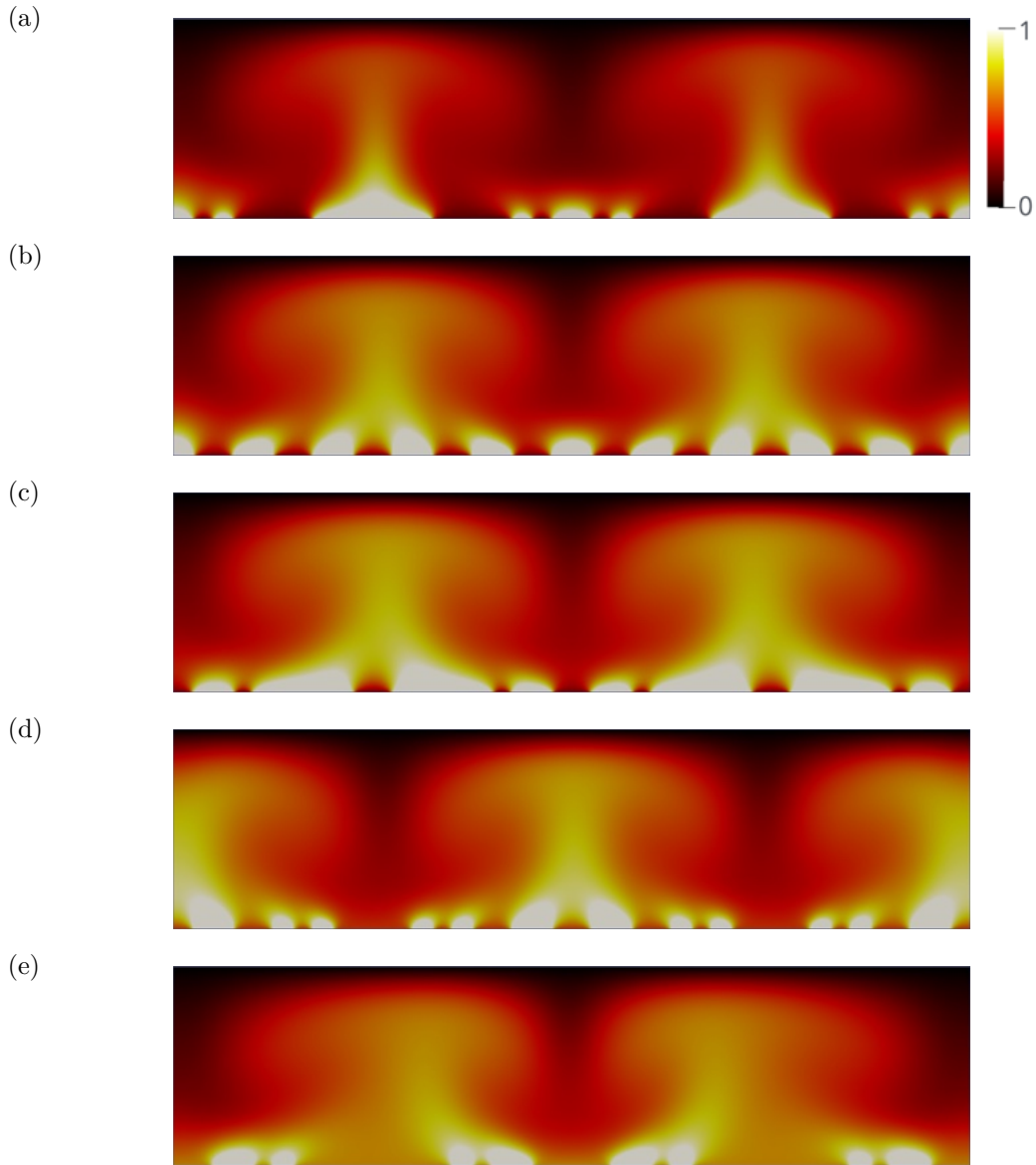


Figure 6: Iso-contours of the steady-state controlled temperature, computed between 0 and 1 under several zero-mean temperature distributed at the hot bottom wall. Each sub-plot illustrates a different learning episode performed over the course of the DRL optimization.

370 number. The authors in Ref. [15] conversely report complete suppression using a classical multi-step  
 371 DRL algorithm adjusting dynamically the temperature from appropriate sensing of flow changes,  
 372 but this only reflects the sub-optimality of operating under an open-loop strategy. For the sake  
 373 of completeness, we note that a mitigation in similar proportions (with efficiency of about 23%)  
 374 is reported in Ref. [17] using multi-step DRL, but this is yet another setup in which symmetry is  
 375 assumed at the lateral ends of the cavity, which makes it difficult to further compare.

## 376 5. Control of quenching cooling rate in a 2-D open tank

### 377 5.1. General case description

378 We aim now at controlling quenching in the two-dimensional, rectangular tank described in  
 379 figure 10, that has width 0.6m and height 0.4m. Four workpiece geometries of increasing complexity  
 380 are considered, as seen in figure 11 that we refer to as rectangular brick, U-bend, serpentine and  
 381 teeth, and whose width  $d$  (that drives the immersion depth to allow adjusting the insertion angle



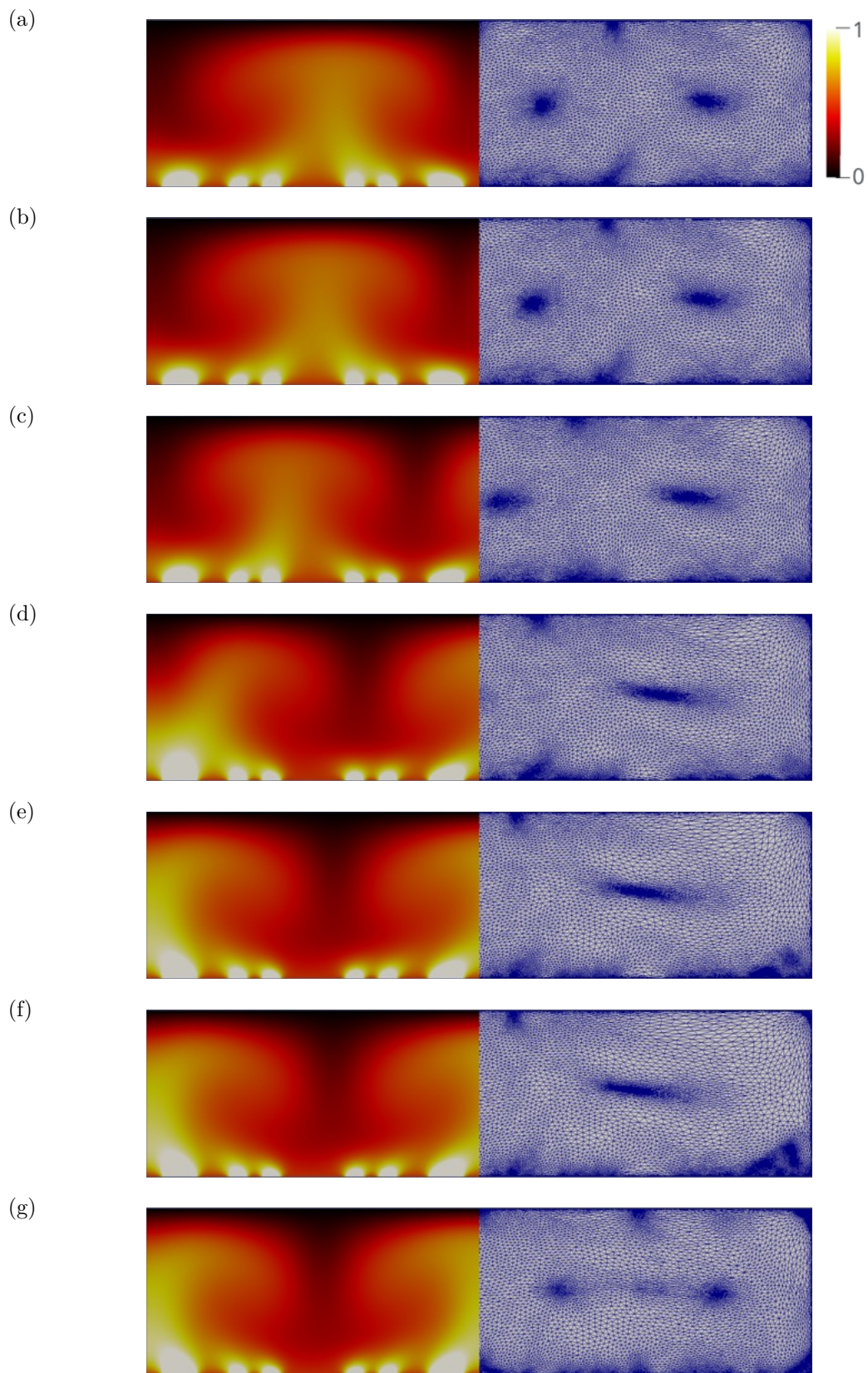


Figure 7: Instantaneous distribution of the controlled temperature, computed between 0 and 1 for the learning episode leading to the steady state in figure 6(d). The snapshots are sampled (from top to bottom) every 20 time units from 0 to 140 time units, after which the steady state from figure 6(d) is recovered in (g).

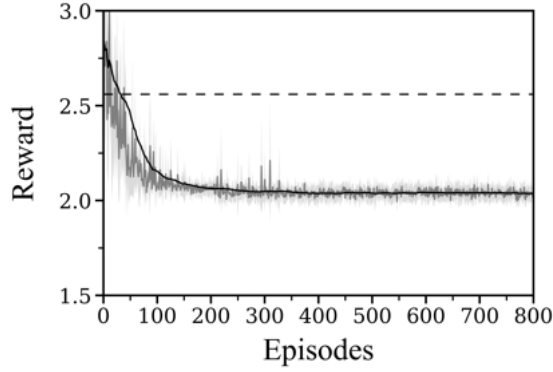


Figure 8: Evolution per learning episode of the instant (in grey) and moving average (in black) Nusselt number, computed by averaging over 20 independent runs. The horizontal dashed line marks the baseline uncontrolled value.

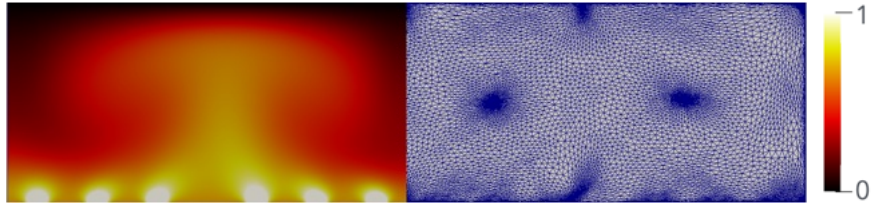


Figure 9: Same as figure 7 for the optimal controlled solution.

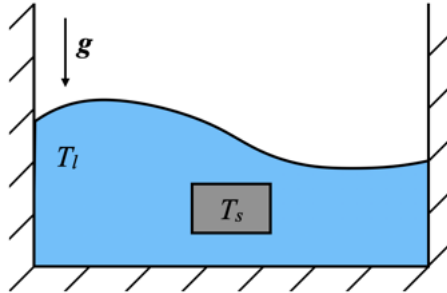


Figure 10: Sketch of the 2-D quenching numerical experiment

382 without contact with the bottom of the tank) is between 0.08m and 0.11m. A Cartesian coordinate  
 383 system is used with origin at the center of the tank, horizontal  $x$ -axis, and vertical  $y$ -axis. The  
 384 quenchant (here, water) is initially at  $T_l = 25^\circ\text{C}$ , while the initial temperature of the metal alloy  
 385 (Inconel) is  $T_s = 880^\circ\text{C}$ . The saturation temperature of water at normal temperature and pressure  
 386 conditions is taken to be  $T_{sat} = 100^\circ\text{C}$ . Dirichlet boundary conditions for velocity and temperature  
 387 boundary conditions are applied on the boundaries of the tank while the top is kept as a free surface.  
 388 Each simulation runs for 60s of cooling time, with time steps  $\Delta t$  ranging between 0.005 and 0.1s  
 389 and number of elements ranging between 20000 and 70000, depending upon the case stiffness. The  
 390 mesh adaptation algorithm presented in section 2 continuously tracks the evolving vapor phase by  
 391 remeshing every 5 time steps with minimum mesh size kept at 0.5mm to minimize errors at the  
 392 interface.

### 393 5.2. Control, reward and sensor placement

394 The quantity being optimized is the orientation of the workpiece measured by its angle  $\theta$  with  
 395 the horizontal, with the understanding that  $\theta$  is positive for a clockwise rotation up to  $180^\circ$ , and  
 396  $\theta = 0^\circ$  corresponds to the baseline orientation shown in figure 11. It is worth emphasizing that  
 397 unlike the above Rayleigh-Bénard convection case, the action space here is continuous, although



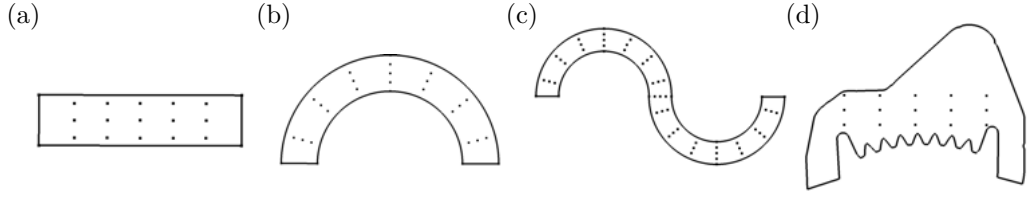


Figure 11: Workpiece geometries for the various test cases in section 5 together with associated sensor placement. (a) Rectangular brick, (b) U-bend, (c) serpentine, and (d) teeth geometry.

$\gamma$	$T_l$	$T_s$	$\mu$	$\rho$	$\lambda$	$c_p$		$d$	$n_p$
							Rectangular brick	0.08	15
							U-bend	0.08	21
							Serpentine	0.11	45
							Teeth	0.08	15
	25	-	0.005	1000	0.6	4185	Liquid		
0.07			0.001	1.7	0.025	2010	Vapor		

Table 1: Numerical parameters used in the 2-D quenching numerical experiment. All values in SI units, with the exception of temperatures given in Celsius.

398 all angles reported in the following are rounded to the nearest integer to ease the reading. The  
 399 network action output therefore consists of a single value  $\hat{x}$  in  $[-1; 1]$ , mapped into the actual angle  
 400 according to

$$\theta = \theta_{max} \frac{\hat{x} + 1}{2} - \theta_{min} \frac{\hat{x} - 1}{2}, \quad (22)$$

401 where we set  $\theta_{max} = -\theta_{min} = 180^\circ$ . An ideal control would target a very small temperature  
 402 difference (ideally zero) between any two points of the workpiece at any given time during the entire  
 403 quenching duration. Since the workpiece in quenching is initially heated to a uniform temperature,  
 404 one way of mathematically achieving this objective is by forming a reward function that penalizes  
 405 the agent proportional to the maximum heat flux flowing out of the workpiece at every instant.  
 406 This forces the DRL agent to homogenize the heat flux out of the workpiece, which achieves  
 407 indirectly homogeneous temperature distribution. Local temperature and temperature gradient  
 408 information (as heat flux is directly proportional to temperature gradient) is thus recorded during  
 409 the simulation at  $N_p$  specific sensor locations in the workpiece, after which we compute the reward  
 410 following the gradient strategy presented in [16], meant to approximate the averaged magnitude  
 411 of the tangential heat flux from (except that all averages are performed here in space and time to  
 412 encompass the whole history of phase change occurring of the quenching process). This information  
 413 can help the agent update its policy so that future actions to minimize the largest gradient in a  
 414 specific direction.

415 The probes information and placement for each geometry are documented in Table 1 and figure  
 416 11, respectively. For the rectangular brick and the teeth geometries, the probes are arranged in  
 417 an array of  $n_x$  columns and  $n_y$  rows with resolutions  $\Delta_x$  and  $\Delta_y$ . The following formula gives an  
 418 estimate of the tangential heat flux by averaging the norm of the temperature gradient in time and  
 419 across rows and columns in  $x$  and  $y$  directions, respectively:

$$\langle \|\nabla_{\parallel} T\| \rangle_i = \frac{2}{n_y - 1} \left| \sum_{j \neq 0} \text{sgn}(j) \|\nabla T\|_{ij} \right|, \quad (23)$$

$$\langle \|\nabla_{\parallel} T\| \rangle_j = \frac{2}{n_x - 1} \left| \sum_{i \neq 0} \text{sgn}(i) \|\nabla T\|_{ij} \right|, \quad (24)$$

420 where subscripts  $i$ ,  $j$  and  $ij$  denote quantities evaluated at  $x = i\Delta_x$ ,  $y = j\Delta_y$  and  $(x, y) =$   
 421  $(i\Delta_x, j\Delta_y)$ , respectively, and symmetrical numbering is used for the center probe to sit at the

422 intersection of the zero-th column and row. The reward  $r_t = -\langle \|\nabla_{\parallel} T\| \rangle$  fed to the DRL agent is  
 423 given by the average of the quantities calculated before as,

$$\langle \|\nabla_{\parallel} T\| \rangle = \frac{d}{T_{ref}} \frac{1}{n_x + n_y} \sum_{i,j} \langle \|\nabla_{\parallel} T\| \rangle_i + \langle \|\nabla_{\parallel} T\| \rangle_j, \quad (25)$$

424 which specially yields  $r_t = 0$  for a perfectly homogeneous cooling. For the U-bend and serpentine  
 425 geometries, a conformal mapping from rectangular to circular sector is used to distribute a similar  
 426 array of probes on the actual workpiece geometry, in which case the above relations carry over  
 427 provided the  $i$  and  $j$  indices are taken to refer to the *unmapped* sensor positions.

### 428 5.3. Rectangular brick test case

429 We start with the most straightforward and popular quenching geometry possible: the rectan-  
 430 gular brick. The evolution of the rewards and actions for 320 episodes (40 episodes) is presented  
 431 in figure 12(a) and (b), respectively. Initially, the agent randomly explores the action space to  
 432 learn from the rewards. It is rather intuitive that for  $\theta = 0^\circ$  configuration shown in figure 13(a),  
 433 the uneven vapor accumulation will cause differential cooling at the top and bottom surface of  
 434 the workpiece, in turn leading to differential properties and thermal distortions. Moreover, the  
 435 associated vapor film evolution pattern shows that a thick vapor film forms on the upper surface  
 436 compared to the lower surface, that creates thermal insulation. For these reasons, it is standard  
 437 practice in the industry to avoid this configuration. Nonetheless, it is interesting that the DRL  
 438 agent learns this aspect without prior knowledge of boiling physics in the first few episodes. As  
 439 the insertion angle increases to  $\theta = 17^\circ$  (rounded to the closest integer), the vapor film insulation  
 440 accumulates on the top left corner, causing high temperature in this corner up to  $t=20-30s$ ; see  
 441 figure 13(b). It is worth mentioning that the workpiece temperature is close to the recrystalliza-  
 442 tion temperature in some zones; hence such skewed temperature distribution will cause worst grain  
 443 growth (and material properties) than in the previous case. The reward keeps decreasing until  
 444 the brick is again set to achieve similar states due to the geometrical symmetry, after which the  
 445 DRL agent settles in a range from  $50^\circ$  to  $150^\circ$  (which is rather fortuitous since the agent does  
 446 not learn about symmetries under the optimization process 16), and it takes a dozen episodes  
 447 for the variance to start diminishing. At this stage, the insertion angle  $\theta = 68^\circ$  shown in figure  
 448 13(c) yields an overall better temperature evolution throughout the quenching process (compared  
 449 to the early actions taken by the agent). It is intriguing to see that for this configuration, the vapor  
 450 film does not initially accumulate on either of the large edges, and the bubble nucleation occurs  
 451 from both surfaces, producing a homogeneous temperature distribution. However, after  $t=30s$ , the  
 452 bubble nucleation stops on the left side, creating a thin vapor film, after which the temperature  
 453 distribution is skewed for the rest of the process. It is out of the scope of this work to analyze  
 454 in details this subtle bubble dynamics, but these observations suffice to highlight the stiffness of  
 455 boiling physics applied to quenching at different insertion angles, as it is at least challenging, if not  
 456 impossible, for an experienced professional to anticipate the behaviors discussed herein above just  
 457 by inspection, even in a simplified 2-D case. The optimal insertion angle for this case is found to  
 458 be  $\theta = 97^\circ \pm 2^\circ$ , with the associated reward  $r_t = 235 \pm 17$ . This is slightly tilted from  $\theta = 90^\circ$ ,  
 459 which forces the upper surface to convect more vapor mass due to the latter having lower specific  
 460 density. It is also shown in figure 13(d) to avoid vapor entrapment along the long and short edges  
 461 while attaining maximum effective length to improve natural convection. It is imperative to note  
 462 that, in this case, the bubble nucleation occurs throughout the process with equal intensity from  
 463 both surfaces. This can be one of the indications (along with the uniform temperature profiles)  
 464 that the current state is arguably the optimum state achieved by the DRL.

### 465 5.4. U-bend test case

466 The U-bend is another common geometries treated in the industry. Similar to the previous case,  
 467 the test case is set up with a total of 320 simulations (40 episodes), whose rewards and actions  
 468 evolution is shown in figure 16. One of the most popular configurations for this geometry is the in  
 469 the quenching industry is the inverted-U shape 38 illustrated in figure 15(a) that corresponds to  
 470  $\theta = 180^\circ$ . In this case, the nucleation of the bubbles mostly occurs at the top of the curved part  
 471 while the two thongs are covered in a thick vapor film. This leads to a differential temperature

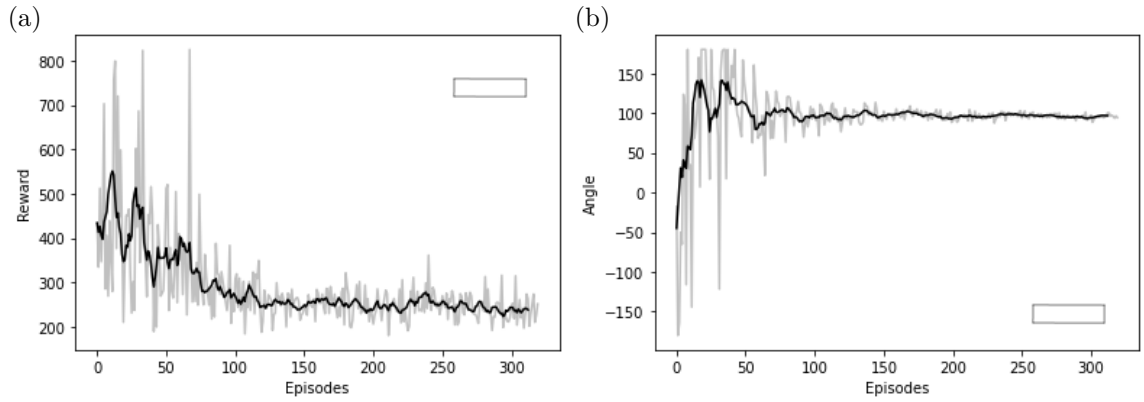


Figure 12: (a) Evolution per learning episode of the instant (in grey) and moving average (in black) reward for the rectangular brick test case. (b) Same as (a) for the workpiece insertion angle.



Figure 13: Temperature evolution for the rectangular brick test case inserted at various angles. (a)  $\theta = 0^\circ$ , (b-d) random angles selected over the course of optimization, and (e)  $\theta = 97^\circ$ , the optimal angle selected by the DRL agent. The snapshots are sampled (from left to right) every 10s from 0 to 50s. The iso-contours in the solid (resp. in the fluid) show the temperature field between 0 and  $880^\circ\text{C}$  (resp. the density field between 1.7 and  $8000\text{ kg/m}^3$ ).

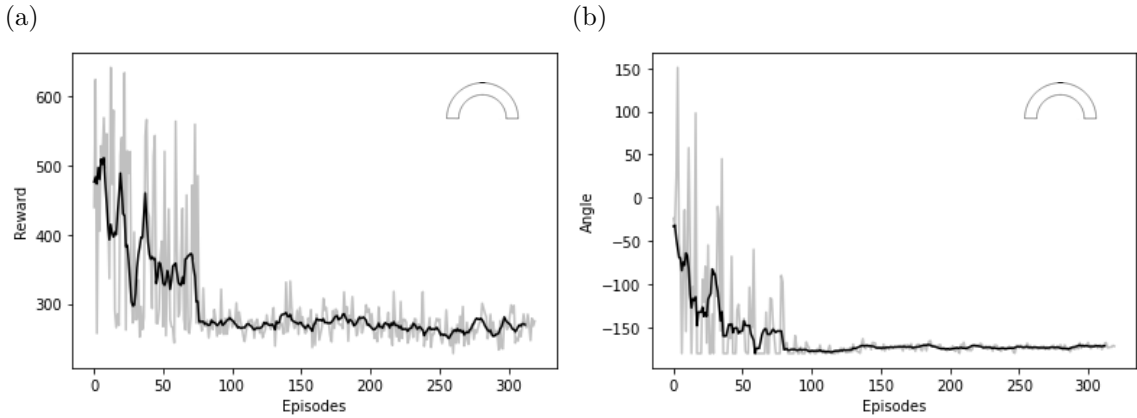


Figure 14: (a) Evolution per learning episode of the instant (in grey) and moving average (in black) reward for the U-bend test case. (b) Same as (a) for the workpiece insertion angle.

472 distribution within the solid as the mid portion of the workpiece cools faster than two thongs  
 473 creating three different temperature zones. Another obvious strategy is the U configuration in  
 474 figure 15(b), that corresponds to  $\theta = 0^\circ$ . It yields a more uniform temperature distribution than  
 475 the inverted U configuration, but one that i remains skewed. Meanwhile, it takes DRL agent  
 476 approximately 50 episodes to find an optimal insertion angle  $\theta = -171^\circ \pm 2^\circ$  associated to reward  
 477  $269 \pm 8$ . The latter corresponds to a beneficial tilt of the U configuration that slightly reduces the  
 478 skewness in the temperature profile, and achieves a more homogeneous distribution at all profile  
 479 at all intermediate time steps, as shown in figure 15(d). Again, we only aim here at assessing  
 480 the ability to output unanticipated quenching solutions using DRL, so explaining the complex  
 481 underlying physics remains out of scope at this stage, but it should be noted again that such  
 482 dynamics are complicated to spot with mere observation, even to the shrewd eye.

### 483 5.5. Serpentine

484 The serpentine (or double U-bend) is a more complex geometry for which inferring an optimal  
 485 insertion angle is a really challenging task. We train here the DRL agent with a total of 640  
 486 simulations (80 episodes), with rewards and actions evolution reported in figure 16. For this case,  
 487 it takes about 25 episodes to converge to an optimal  $77^\circ \pm 2^\circ$ , which yields a reward  $863 \pm 20$ . In  
 488 this setting, the temperature profile in the two U-bends is almost the same (symmetric) over the  
 489 whole quenching process, as shown in figure 17(d). This is highly non-trivial, given the difficulty  
 490 to achieve homogeneous cooling in a single U-bend, discussed in earlier sections. By comparison,  
 491 at  $\theta = 0^\circ$  (figure 17(a)), one half of the geometry loses heat faster than the other one (which can  
 492 yield increased residual stresses), while at  $\theta = 90^\circ$  (figure 17(b)), the upper and middle parts of the  
 493 geometry remain hotter compared to the other sections. The other angle reported in figure 17(c)  
 494 yield differential cooling in the upper and middle section as the vapor gets entrapped in this region,  
 495 insulating it further.

### 496 5.6. Teeth geometry

497 The teeth geometry is a highly complicated geometry inspired by actual industrial components,  
 498 one that has no axis of symmetry, which makes inferring an optimal insertion angle completely  
 499 impossible. The agent is trained with a total of 320 simulations (40 episodes); see figure 18 for  
 500 the associated rewards and actions. While we do not strictly speaking achieve convergence for this  
 501 case, as evidenced by the substantial variations in the insertion angles achieved over the last part of  
 502 training, the agent succeeds in identifying a relevant range of parameters in the order from 100 to  
 503  $180^\circ$ . If we compare to the results obtained at empiric angles ( $\theta = 0^\circ, 45^\circ, 90^\circ$ ), it is seen in figure 19  
 504 that DRL produces a rather homogeneous temperature profile. As evident from previous cases, the  
 505 workpiece face where vapor film accumulates and bubble creation occurs is always hotter compared  
 506 to other regions. At  $\theta = 90^\circ$ , this vapor accumulation causes skewed temperature profile in both  
 507 horizontal (due to vapor accumulation) and vertical (due to thickness) directions; see figure 19(a).  
 508 In other words, the geometry region that is thinner and immersed deeper, always cools first. This

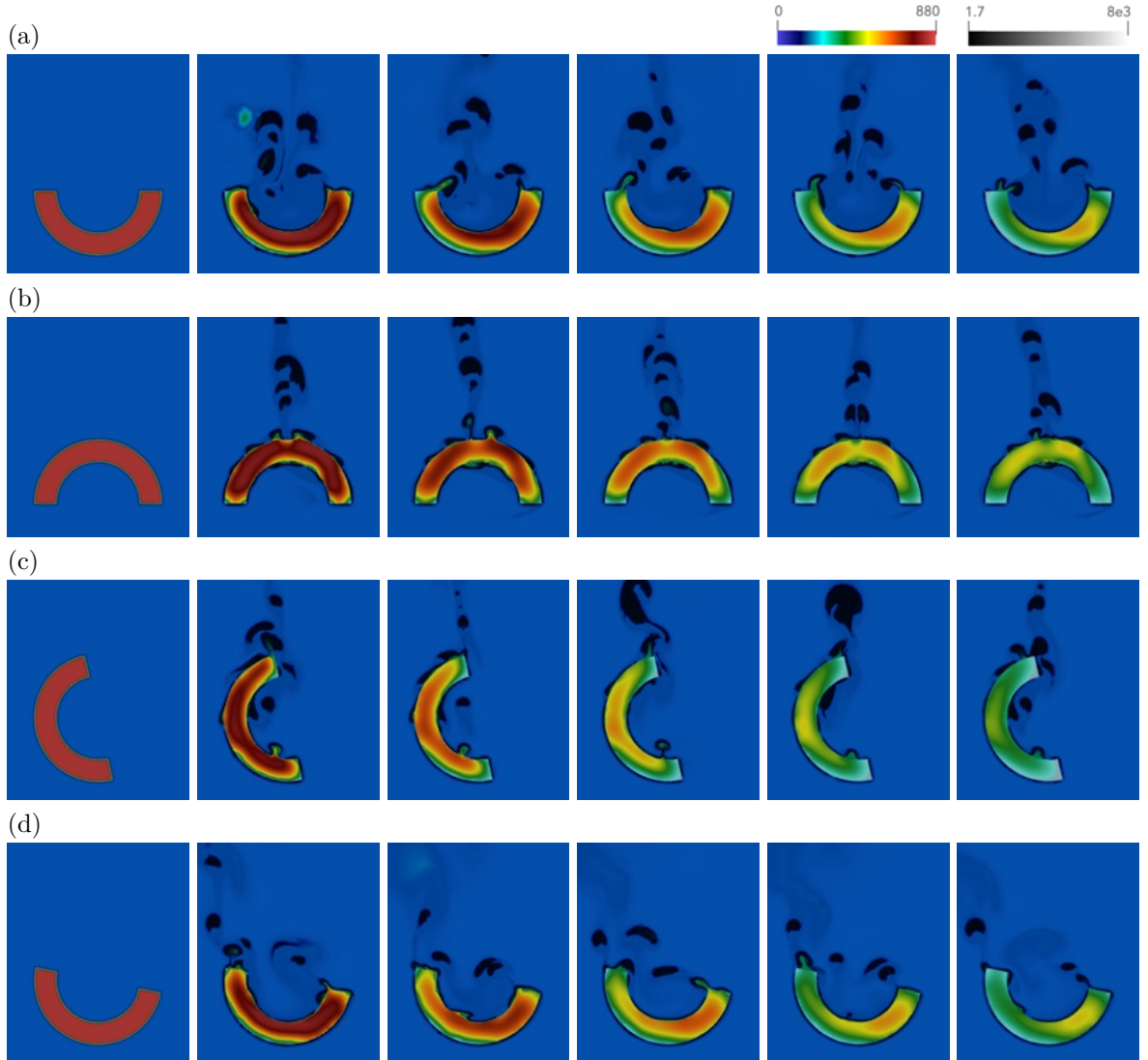


Figure 15: Temperature evolution for the U-bend test case inserted at various angles. (a)  $\theta = 0^\circ$ , (b)  $\theta = 180^\circ$ , (c)  $\theta = -85^\circ$ , and (d)  $\theta = -171^\circ$ , the optimal angle selected by the DRL agent. The snapshots are sampled (from left to right) every 10s from 0 to 50s. The iso-contours in the solid (resp. in the fluid) show the temperature field between 0 and  $880^\circ\text{C}$  (resp. the density field between 1.7 and  $8000\text{ kg/m}^3$ ).

509 carries over to the other empiric setting considered, namely  $\theta = 0^\circ$  in figure 19(b) and  $\theta = 45^\circ$  in  
 510 figure 19(c), where the thick region stays hotter compared to the teethes and thongs. Despite the  
 511 lack of convergence (discussed in the next section), the DRL agent seems to learn these nuances  
 512 by proposing a solution to insert the workpiece upside down, which increases the heat flux from  
 513 the thick region (due to natural convection), while the heat flux from the thin teethed decreases  
 514 due to added insulation coming from the vapor films.

## 515 6. Conclusion and recommendations for future research

### 516 6.1. Conclusion

517 In this work, a numerical framework is presented, in which a fully connected network learns  
 518 to find optimal parameters in the process of quenching heat treatment. The agent is trained  
 519 with the single-step PPO deep reinforcement algorithm, and gets only one attempt per learning  
 520 episode at finding the optimal. The numerical reward fed to the network is computed with a

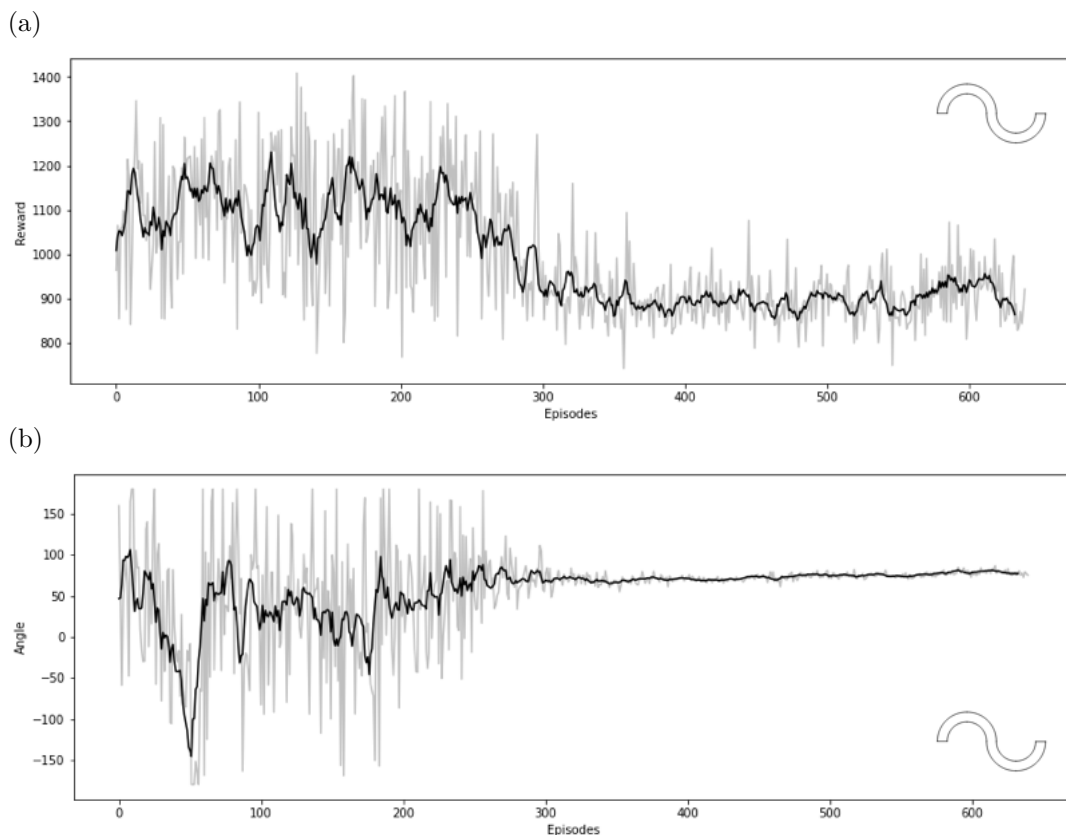


Figure 16: (a) Evolution per learning episode of the instant (in grey) and moving average (in black) reward for the serpentine test case. (b) Same as (a) for the workpiece insertion angle.

521 stabilized finite elements CFD environment solving a phase change model formulated after pseudo-  
 522 compressible Navier–Stokes and heat equations, using a combination of variational multi-scale  
 523 modeling, immerse volume method, and multi-component anisotropic mesh adaptation.

524 Relevance of the proposed methodology is illustrated by controlling natural convection in a  
 525 closed cavity with aspect ratio 4:1, for which DRL alleviates the flow-induced enhancement of heat  
 526 transfer by approximately 20%. Regarding quenching applications, the DRL algorithm succeeds in  
 527 finding optimal orientations that adequately homogenize the temperature distribution within both  
 528 simple and complex 2-D part geometries, and improve over simpler trial-and-error configurations  
 529 classically used in the quenching industry. Such results clearly stress that single-step PPO (and  
 530 DRL in general) can be effective to explore and discover new solutions from unforeseen parameter  
 531 combinations in quenching applications.

## 532 6.2. Strategies towards practical application

533 As an exploratory study, the current research provides preliminary evidence of the ability of the  
 534 proposed DRL-CFD framework in optimizing complex quenching processes by improving cooling  
 535 uniformity to mitigate thermal residual stresses. The presented results are encouraging, but more  
 536 work is needed to to confirm and extend our conclusions, and to fully scope out the potential of the  
 537 approach in real-world scenarios. In concluding the present paper, it is thus proposed to discuss  
 538 key directions for improvement, all intended to help bridge the gap between the current capabilities  
 539 and the requirements of practical deployment.

540 Quenching is a complex thermomechanical process that can be cast as a thermal fluid-structure  
 541 interaction problem involving the simultaneous resolution of turbulent flows with phase change  
 542 and conjugate heat transfers between the solid and the fluid subdomains. Overall, the field is ever  
 543 evolving, and there is a clear need for improved CFD models capable of dealing with this problem in  
 544 all its complexity. A high-fidelity adaptive, multiphase DRL-CFD framework predicting accurately  
 545 the phase change at the liquid-vapor interface, but also the phase transformation of the treated



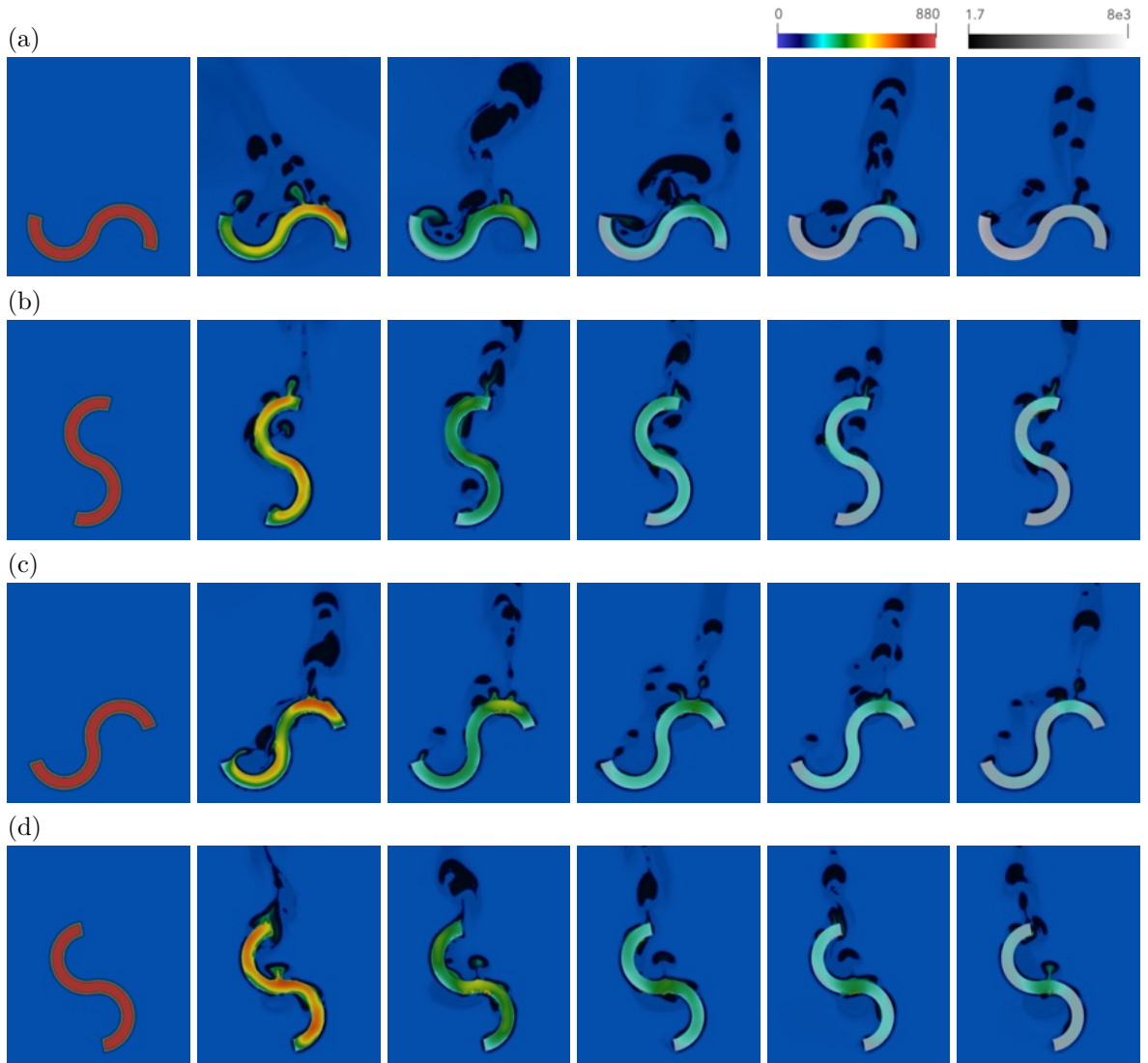


Figure 17: Temperature evolution for the serpentine test case inserted at various angles. (a)  $\theta = 0^\circ$ , (b)  $\theta = 90^\circ$ , (c) random angle selected over the course of optimization, and (d)  $\theta = 77^\circ$ , the optimal angle selected by the DRL agent. The snapshots are sampled (from left to right) every 10s from 0 to 50s. The iso-contours in the solid (resp. in the fluid) show the temperature field between 0 and  $880^\circ\text{C}$  (resp. the density field between 1.7 and  $8000\text{ kg/m}^3$ ).

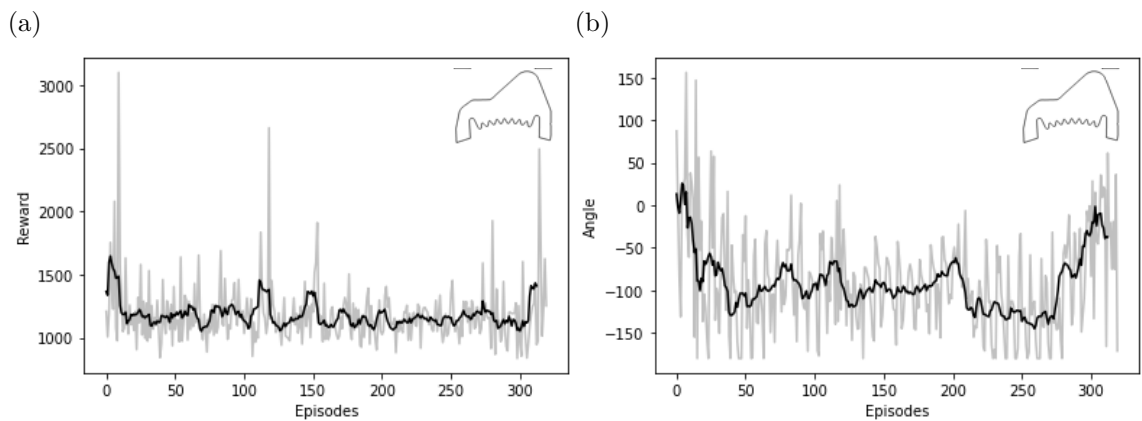


Figure 18: (a) Evolution per learning episode of the instant (in grey) and moving average (in black) reward for the teeth geometry test case. (b) Same as (a) for the workpiece insertion angle.

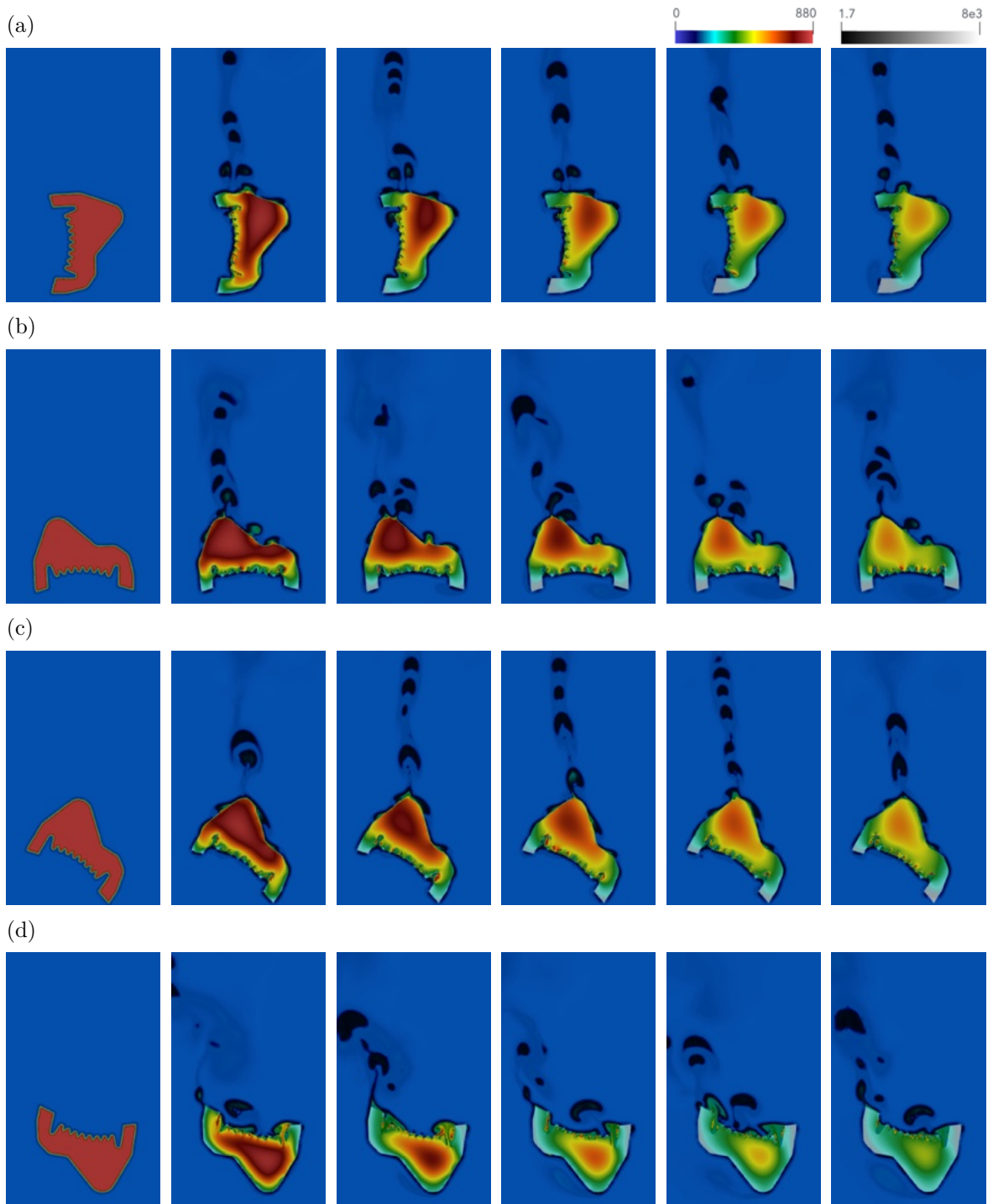


Figure 19: Temperature evolution for the teeth geometry test case inserted at various angles. (a)  $\theta = 90^\circ$ , (b)  $\theta = 0^\circ$ , (c)  $\theta = 45^\circ$ , and (d)  $\theta = 180^\circ$ , in the optimal range selected by the DRL agent. The snapshots are sampled (from left to right) every 10s from 0 to 50s. The iso-contours in the solid (resp. in the fluid) show the temperature field between 0 and 880°C (resp. the density field between 1.7 and 8000 kg/m<sup>3</sup>).

546 part to predict its metallurgical evolution, will thus be instrumental in providing industrially rel-  
 547 evant process parameters encompassing not only the thermal residual stresses caused by large  
 548 temperature gradients (as has been done here), but also the volumetric residual stresses caused by  
 549 martensitic transformations. By then, it is reasonable to expect that further developments in the  
 550 fast-moving field of deep reinforcement learning will allow for faster convergence and lesser execu-

tion load (using, *e.g.*, auto-encoders and systematic state compression, or on-the-fly generation of surrogate models with uncertainty level prediction), and will facilitate application to industrially relevant 3-D configurations. This should set up a framework fast enough to inform process design in a matter of hours rather than days, thereby reliably augmenting industrial applicability. A core feature of the proposed framework in this respect is its high generalizability, that is, the fact that it builds naturally from any improved CFD model or/and DRL training method. For instance, a more elaborated algorithm can easily substitute for the rather simplistic PPO framework, such as Policy-based Optimization [36], another single-step reinforcement algorithm that samples actions from full covariance matrices, and is theoretically better suited to represent higher order logic and to handle complex parameter interactions.

An important limiting factor is the limited availability of process data, often affected to a small number of sensors, which can make it difficult to develop accurate models and control algorithms due to high-variance output. In a broad sense, the agent operates under partially observable environments, so its performance is highly dependent on the quality and relevance of the available data (this is an issue strongly related to data-driven model reduction techniques for large scale dynamical systems, which usually require using measures of observability as an information quality metric). In this regards, we note that the reward construction strategy employed in this work puts constraints on how many sensors can be used and how they can be arranged in the workpiece. This is detrimental to learning, and may explained the overall lack of convergence observed in the teeth geometry, as the agent does not learn the heat flux out of the workpiece in some regions of interest (here, the hood) but has to figure it out from the indirect information it gets from the other sensors. Another explanation is that the reward is approximated from point-wise temperature data (similar to experimental measurements) that has more sensitivity to small numerical errors (*e.g.*, the interpolation error at the probes position) than an integral quantity, and mesh adaptation procedure is not a deterministic process, as the outcome depends on the processors and number of processors used, and any initial difference propagates over the course of the simulation because the meshes keep being adapted dynamically. For these reasons, two control parameters, even close, can yield different rewards on behalf of different interpolation errors at the probes position and different nucleation patterns initiated by slightly different initial conditions, as illustrated in figures 20 for the rectangular brick, the U-bend and the serpentine solution. This also likely explains why the variance in reward is systematically larger (by a factor of almost 5) than that of the action itself for all cases reported herein. Ultimately, it calls for the design of robust reward functions capable of guiding the learning process toward effective and efficient policies even with randomly distributed sensors. This is no small task, in the absence of a best practice on how to design a reward function (this being essentially a trial-and-error process of a practitioner using their knowledge to define a baseline reward intended to provide a consistent feedback to the agent about its performance, observing how the agent performs, then tweaking the reward to achieve greater performance).

Finally, another reason to push DRL forward in this context is the ability of neural networks to transfer knowledge from previous experiences, to quickly adapt to different environments (workpiece geometry and material properties, quenchant) and effectively learn new tasks. For instance, it is easy to compare different settings of design complexity, reflecting different levels of constrained operation when it comes to optimizing a practically meaningful scenarios (*e.g.*, heavily constrained optimization problems relevant to cases where the practitioner has limited freedom to optimize the design, in which case one can seek to optimize the orientation, immersion rate and depth of the solid part and the fluid viscosity, or mildly constrained problems relevant to cases where the practitioner has great freedom to act, in which case additional parameters can include the size of the tank, the number, type and placement of agitators and the agitation rate). We expect that this will be a key feature to reduce learning time and improve neural network performance, as progress are made towards realizing the potential of DRL-CFD for flexible, ready-to-use control of industrial manufacturing processes.

## References

- [1] Z. Zhao, M. Stuebner, J. Lua, N. Phan, J. Yan, Full-field temperature recovery during water quenching processes via physics-informed machine learning, *Journal of Materials Processing Technology* 303 (2022) 117534.

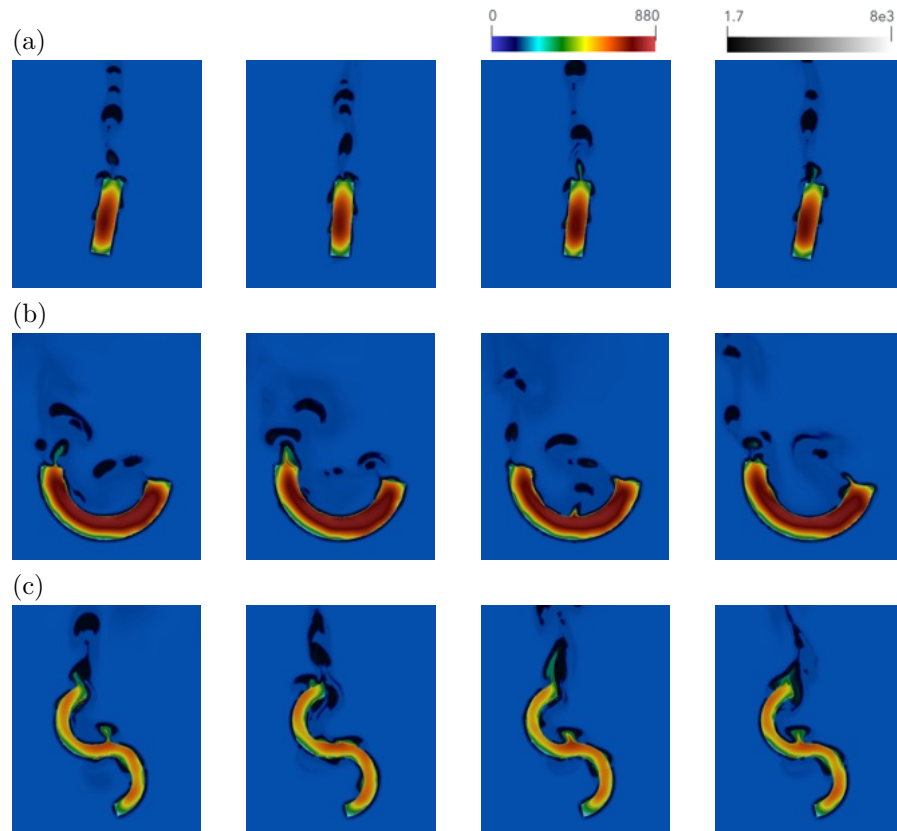


Figure 20: Temperature at  $t = 20$ s for the (a) rectangular brick, (b) U-bend, and (c) serpentine geometries inserted at several angles close to the DRL optimal.

- 605 [2] R. D. Lopez-Garcia, I. Medina-Juárez, A. Maldonado-Reyes, Effect of quenching parameters  
606 on distortion phenomena in AISI 4340 steel, *Metals* 12 (5) (2022).
- 607 [3] L. He, H. Li, Fem simulation of quenching residual stress for the plane strain problems, in:  
608 2010 International Conference On Computer Design and Applications, Vol. 3, 2010, pp. V3-  
609 119-V3-123.
- 610 [4] M. Decroos, M. Seefeldt, The effect of size on the distortion behavior after carburisation and  
611 quenching processes of gears, *Int. J. Met. Mater. Eng* 139 (2017) 1–10.
- 612 [5] S. Šolić, B. Podgornik, V. Leskovšek, The occurrence of quenching cracks in high-carbon  
613 tool steel depending on the austenitizing temperature, *Engineering failure analysis* 92 (2018)  
614 140–148.
- 615 [6] A. da Silva, T. Pedrosa, J. Gonzalez-Mendez, X. Jiang, P. Cetlin, T. Altan, Distortion in  
616 quenching an AISI 4140 C-ring - Predictions and experiments, *Materials & Design* 42 (2012)  
617 55–61.
- 618 [7] L. Pinto, M. Andrychowicz, P. Welinder, W. Zaremba, P. Abbeel, Asymmetric actor critic for  
619 image-based robot learning, arXiv preprint arXiv:1710.06542 (2017).
- 620 [8] D. Bahdanau, P. Brakel, K. Xu, A. Goyal, R. Lowe, J. Pineau, A. Courville, Y. Bengio, An  
621 actor-critic algorithm for sequence prediction, arXiv preprint arXiv:1607.07086 (2016).
- 622 [9] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, M. Riedmiller,  
623 Playing Atari with Deep Reinforcement Learning, arXiv preprint arXiv:1312.5602 (2013).
- 624 [10] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert,  
625 L. Baker, M. Lai, A. Bolton, Y. Chen, T. Lillicrap, F. Hui, L. Sifre, G. van den Driessche,

- 626 T. Graepel, D. Hassabis, Mastering the game of go without human knowledge, *Nature* 550  
627 (2017) 354–359.
- 628 [11] A. Kendall, J. Hawke, D. Janz, P. Mazur, D. Reda, J.-M. Allen, V.-D. Lam, A. Bewley,  
629 A. Shah, Learning to drive in a day, arXiv preprint arXiv:1807.00412 (2018).
- 630 [12] W. Knight, Google just gave control over data center cooling to an AI, *MIT Technology*  
631 *Review* (2018).
- 632 [13] P. Garnier, J. Viquerat, J. Rabault, A. Larcher, A. Kuhnle, E. Hachem, A review on deep  
633 reinforcement learning for fluid mechanics, *Comp. Fluids* 225 (2021) 104973.
- 634 [14] J. Viquerat, P. Meliga, A. Larcher, E. Hachem, A review on deep reinforcement learning for  
635 fluid mechanics : an update, *Phys. Fluids* 34 (2022) 111301.
- 636 [15] G. Beintema, A. Corbetta, L. Biferale, F. Toschi, Controlling Rayleigh–Bénard convection via  
637 reinforcement learning, *Journal of Turbulence* (2020) 585–605.
- 638 [16] E. Hachem, H. Ghraieb, J. Viquerat, A. Larcher, P. Meliga, Deep reinforcement learning for  
639 the control of conjugate heat transfer, *J. Comput. Phys.* 436 (2021) 110317.
- 640 [17] C. Vignon, J. Rabault, J. Vasanth, F. Alcántara-Ávila, M. Mortensen, R. Vinuesa, Effective  
641 control of two-dimensional Rayleigh–Bénard convection: invariant multi-agent reinforcement  
642 learning is all you need, *Phys. Fluids* 36 (2023) 065146.
- 643 [18] M. Renault, J. Viquerat, P. Meliga, G.-A. Grandin, N. Meynet, E. Hachem, Investigating gas  
644 furnace control practices with reinforcement learning, *Int. J. Heat Mass Transfer* 209 (2023)  
645 124147.
- 646 [19] Y.-Z. Wang, J.-Z. Peng, N. Aubry, Y.-B. Li, Z.-H. Chen, W.-T. Wu, Control policy transfer  
647 of deep reinforcement learning based intelligent forced heat convection control, *Int. J. Therm.*  
648 *Sci.* 195 (2024) 108618.
- 649 [20] H. Keramati, F. Hamdullahpur, M. Barzegari, Deep reinforcement learning for heat exchanger  
650 shape optimization, *Int. J. Heat Mass Transfer* 194 (2022) 123112.
- 651 [21] A. di Meglio, N. Massarotti, P. Nithiarasu, A physics-driven and machine learning-based  
652 digital twinning approach to transient thermal systems, *Int. J. Numer. Methods Heat Fluid*  
653 *Flow* ahead-of-print (2024) ahead-of-print.
- 654 [22] T. Zhang, J. Luo, P. Chen, J. Liu, Flow rate control in smart district heating systems using  
655 deep reinforcement learning, arXiv preprint arXiv:1912.05313 (2019).
- 656 [23] M. Kim, J. H. Moon, Deep neural network prediction for effective thermal conductivity and  
657 spreading thermal resistance for flat heat pipe, *Int. J. Numer. Methods Heat Fluid Flow* 33  
658 (2022) 437–455.
- 659 [24] J. Viquerat, J. Rabault, A. Kuhnle, H. Ghraieb, A. Larcher, E. Hachem, Direct shape opti-  
660 mization through deep reinforcement learning, *J. Comput. Phys.* 428 (2021) 110080.
- 661 [25] H. Ghraieb, J. Viquerat, A. Larcher, P. Meliga, E. Hachem, Single-step deep reinforcement  
662 learning for open-loop control of laminar and turbulent flows, *Phys. Rev. Fluids* 6 (2021)  
663 053902.
- 664 [26] M. Khalloufi, Multiphase flows with phase change and boiling in quenching processes, Ph.D.  
665 thesis, PSL Research University (2017).
- 666 [27] M. Khalloufi, R. Valette, E. Hachem, Adaptive eulerian framework for boiling and evaporation,  
667 *J. Comput. Phys.* 401 (2020) 109030.
- 668 [28] E. Hachem, T. Kloczko, H. Dignonnet, T. Coupeuz, Stabilized finite element solution to handle  
669 complex heat and fluid flows in industrial furnaces using the immersed volume method, *Int.*  
670 *J. Numer. Meth. Eng.* 68 (2012) 99–121.

- 671 [29] E. Hachem, H. Digonnet, E. Massoni, T. Coupez, Immersed volume method for solving natural  
672 convection, conduction and radiation of a hat-shaped disk inside a 3d enclosure, *Int. J. Numer.*  
673 *Method H.* 22 (2012) 718–741.
- 674 [30] C. Gruau, T. Coupez, 3D tetrahedral, unstructured and anisotropic mesh generation with  
675 adaptation to natural and multidomain metric, *Comput. Methods Appl. Mech. Engrg.* 194  
676 (2005) 4951–4976.
- 677 [31] M. M. Bernitsas, K. Raghavan, Y. Ben-Simon, E. M. H. Garcia, Vivace (vortex induced  
678 vibration aquatic clean energy): a new concept in generation of clean and renewable energy  
679 from fluid flow, *J. Offshore Mech. Arctic Engrg.* 130 (2008) 041101.
- 680 [32] Y. Mesri, H. Digonnet, T. Coupez, Advanced parallel computing in material forming with  
681 CIMLib, *Eur. J. Comput. Mech.* 18 (2009) 669–694.
- 682 [33] T. Coupez, Metric construction by length distribution tensor and edge based error for  
683 anisotropic adaptive meshing, *J. Comput. Phys.* 230 (2011) 2391–2405.
- 684 [34] J. U. Brackbill, D. B. Kothe, C. Zemach, A continuum method for modeling surface tension,  
685 *Journal of computational physics* 100 (2) (1992) 335–354.
- 686 [35] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, O. Klimov, Proximal Policy Optimization  
687 Algorithms, arXiv preprint arXiv:1707.06347 (2017).
- 688 [36] J. Viquerat, R. Duvigneau, P. Meliga, A. Kuhnle, E. Hachem, Policy-based optimization:  
689 single-step policy gradient method seen as an evolution strategy, *Neural Comput. Appl.* 35  
690 (2023) 449–467.
- 691 [37] K. Lee, S. A. Kim, J. Choi, Deep reinforcement learning in continuous action spaces: a case  
692 study in the game of simulated curling, in: *Procs. of the 35th International Conference on*  
693 *Machine Learning*, 2018, pp. 2937–2946.
- 694 [38] Y. Nakagawa, K.-I. Mori, S. Yashima, T. Kaido, Springback behaviour and quenchability in  
695 hot stamping of thick sheets, *Procedia Manufacturing* 15 (2018) 1071–1078.