



**HAL**  
open science

## Single-step deep reinforcement learning for open-loop control of laminar and turbulent flows

Hassan Ghraieb, Jonathan Viquerat, Aurélien Larcher, P. Meliga, Elie Hachem

► **To cite this version:**

Hassan Ghraieb, Jonathan Viquerat, Aurélien Larcher, P. Meliga, Elie Hachem. Single-step deep reinforcement learning for open-loop control of laminar and turbulent flows. *Physical Review Fluids*, 2021. hal-03027908v2

**HAL Id: hal-03027908**

**<https://hal.science/hal-03027908v2>**

Submitted on 17 Nov 2021

**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.

# Single-step deep reinforcement learning for open-loop control of laminar and turbulent flows

H. Ghraieb,<sup>1</sup> J. Viquerat,<sup>1</sup> A. Larcher,<sup>1</sup> P. Meliga,<sup>1</sup> and E. Hachem<sup>1</sup>

<sup>1</sup>*MINES ParisTech, PSL Research University,  
Centre de mise en forme des matériaux (CEMEF),  
CNRS UMR 7635, 06904 Sophia Antipolis Cedex, France*

(Dated: November 16, 2021)

This research gauges the ability of deep reinforcement learning (DRL) techniques to assist the optimization and control of fluid mechanical systems. It relies on single-step PPO, a novel, “degenerate” version of the proximal policy optimization (PPO) algorithm, intended for situations where the optimal policy to be learnt by a neural network does not depend on state, as is notably the case in open-loop control problems. The numerical reward fed to the neural network is computed with an in-house stabilized finite elements environment implementing the variational multiscale (VMS) method. Several prototypical separated flows in two dimensions are used as testbed. The method is applied first to two relatively simple optimization test cases (maximizing the mean lift of a NACA 0012 airfoil and the fluctuating lift of two side-by-side circular cylinders, both in laminar regimes) to assess convergence and accuracy by comparing to in-house DNS data. The potential of single-step PPO for reliable black-box optimization of computational fluid dynamics (CFD) systems is then showcased by tackling several problems of open-loop control with parameter spaces large enough to dismiss DNS. The approach proves relevant to map the best positions for placement of a small control cylinder in the attempt to reduce drag in laminar and turbulent cylinder flows. All results are consistent with in-house data obtained by the adjoint method, and the drag of a square cylinder at Reynolds numbers in the range of a few thousands is reduced by 30%, which matches well reference experimental data available from literature. The method also successfully reduces the drag of the fluidic pinball, an equilateral triangle arrangement of rotating cylinders immersed in a turbulent stream. Consistently with reference machine learning results from the literature, drag is reduced by almost 60% using a so-called boat tailing actuation made up of a slowly rotating front cylinder and two downstream cylinders rotating in opposite directions so as to reduce the gap flow in between them.

Keywords: Deep Reinforcement Learning; Proximal Policy Optimization; Neural Networks; Computational fluid dynamics; Open-loop flow control; Adjoint method

## I. INTRODUCTION

Flow control, defined as the ability to finesse a flow into a more desired state, is a field of tremendous societal and economical importance. In applications such as ocean shipping or airline traffic, reducing the overall drag by just a few percent while maintaining lift can help reducing fossil fuel consumption and CO<sub>2</sub> emission while saving several billion dollars annually [1]. Many other scenario relevant to fluid mechanical systems call for similarly improved engineering design, e.g., the airline industry is greatly concerned with reducing the structural vibrations and the radiated noise that occur under unsteady flow conditions [2, 3], while microfluidics [4] and combustion [5] both benefit from enhanced mixing (which can be achieved by promoting unsteadiness in some appropriate manner). All such problems fall under the purview of this line of study.

Flow control is often benchmarked in the context of bluff body drag reduction. Numerous strategies have been implemented, either open-loop with passive appendices (e.g., end/splitter plates, small secondary cylinder, flexible tail), or open-loop with actuating devices (e.g., plasma actuation, base bleed, rotation) or closed-loop (e.g. transverse motion, blowing/suction, rotation, all relying on appropriate sensing of flow variables); see the comprehensive surveys of recent developments in [6–14]. Nonetheless, most strategies are trial and error and rely on extensive, costly experimental or numerical campaigns, which has motivated the development of rigorous mathematical formalisms capable of achieving optimal design and control with minimal effort. The adjoint method is one family of such algorithms, that has proven efficient at accurately computing the objective gradient with respect to the control variables in large optimization spaces, and has gained prominence in many applications ranging from atmospheric sciences [15] to aerodynamic design [16–19], by way of fresh developments meant to reshape the linear amplification of flow disturbances [20–25].

Another promising option for selecting optimal subsets of control parameters is to rely on machine learning algorithms running labeled data through several layers of artificial neural network

34 while providing some form of corrective feedback. Neural networks are a family of versatile non-  
 35 parametric tools that can learn how to hierarchically extract informative features from data, and  
 36 have gained traction as effective and efficient computational processors for performing a variety  
 37 of tasks, from exploratory data analysis to qualitative and quantitative predictive modeling. The  
 38 increased affordability of high performance hardware (together with reduced costs for data acqui-  
 39 sition and storage) has indeed allowed leveraging the ever-increasing volume of data generated for  
 40 research and engineering purposes into novel insight and actionable information, which in turn  
 41 has reshaped entire scientific disciplines such as image analysis [26] or robotics [27, 28]. Since  
 42 neural networks have produced most remarkable results when applied to stiff large-scale nonlin-  
 43 ear problems [29], it is only natural to assume that they can successfully tackle the state-space  
 44 models arising from the high-dimensional discretization of partial differential equation systems.  
 45 Machine learning has thus been making rapid inroads in fluid mechanics, with consistent efforts  
 46 aimed at solving the governing equations [30], predicting closure terms in turbulence models [31],  
 47 building reduced-order models [32], controlling flows [33, 34], or performing flow measurements and  
 48 visualization [35–37]; see also [38] for an overview of the current developments in this field.

49 The focus here is on deep reinforcement learning (DRL), an advanced branch of machine learning  
 50 in which deep neural networks learn how to behave in an environment so as to maximize some notion  
 51 of long-term reward (a task compounded by the fact that each action affects both immediate and  
 52 future rewards). Several notable works using DRL in mastering games (e.g., Go, Poker) have stood  
 53 out for attaining super-human level [39, 40], but the approach has also breakthrough potential  
 54 for practical applications such as robotics [41, 42], computer vision [43], self-driving cars [44] or  
 55 finance [45], to name a few. There is also great potential for applying DRL to fluid mechanics,  
 56 for which efforts are ongoing but still at an early stage, with only a handful of pioneering studies  
 57 providing insight into the performance improvements to be delivered in shape optimization [46–48]  
 58 and flow control [49–51]. Nonetheless, sustained commitment from the machine learning community  
 59 has allowed expanding the scope from computationally inexpensive, low-dimensional reductions of  
 60 the underlying fluid dynamics [52–54] to complex Navier–Stokes systems [55, 56]. Proximal policy  
 61 optimization (PPO [41]) has quickly gained momentum as one of the go-to algorithms for this  
 62 purpose, as evidenced by several recent publications assessing relevance for open- and closed-loop  
 63 drag reduction in cylinder flows at Reynolds numbers in the range of a few hundreds [57–60].

64 This research draws on this foundation to further shape the capabilities of PPO (still a newcomer  
 65 despite its data efficiency, simplicity of implementation and reliable performance) for flow control,  
 66 and help narrow the gap between DRL and advanced numerical methods for multiscale, multi-  
 67 physics computational fluid dynamics (CFD). The main novelty is the use of single-step PPO, a  
 68 novel “degenerate” algorithm intended for open-loop control problems, as the optimal policy to  
 69 be learnt is then state-independent, and it may be enough for the neural network to get only one  
 70 attempt per episode at finding the optimal. The objective is twofold: first, to prove feasibility  
 71 using several prototypical separated flows in two dimensions as testbed. Second, to assess con-  
 72 vergence and relevance in the context of turbulent flows at moderately large Reynolds number (in  
 73 the range of a few thousands). This is a topic whose surface is barely scratched by the available  
 74 literature, as our literature review did not reveal any other study considering DRL-based control  
 75 of turbulent flows besides [61], another research effort conducted in the same time frame as the  
 76 present work. Single-step PPO has been speculated to hold a high potential as a reliable black-box  
 77 CFD optimizer [48], but we insist that it lies out of the scope of this paper to provide exhaus-  
 78 tive performance comparison data against state-of-the-art optimization techniques (e.g., evolution  
 79 strategies or genetic algorithms). This would indeed require a tremendous amount of time and  
 80 resources even though the efforts for developing the method remain at an early stage (to the best  
 81 of our knowledge, no study in the literature has considered using DRL in a similar fashion) and  
 82 new algorithms cannot be expected to reach right away the level of performance of their more  
 83 established counterparts.

84 The organization is as follows: section II introduces single-step PPO (together with the baseline  
 85 principles of DRL and PPO), and outlines the main features of the finite element CFD environment  
 86 used to compute the numerical reward fed to the neural network. Two simple lift optimization  
 87 problems are presented in section III to assess convergence and accuracy by comparing against in-  
 88 house DNS data. In section IV, the method is applied to two open-loop drag reduction problems  
 89 whose parameter spaces are large enough to dismiss DNS, namely the placement of a small control  
 90 cylinder (for which results computed under laminar and turbulent conditions are compared to  
 91 in-house data obtained by the adjoint method), and the cylinder rotation of a turbulent fluidic

92 pinball. Finally, in section V, the method is thoroughly compared (in terms of scope, applicability  
 93 and performances) to the adjoint method. Evolutionary strategies are also briefly reviewed to put  
 94 our contribution in perspective and discuss the advantages that may be expected once single-step  
 95 PPO is finely tuned and characterized.

## 96 II. METHODOLOGY

### 97 A. Deep reinforcement learning

98 Reinforcement learning (RL) provides a consistent framework for modeling and solving decision-  
 99 making problems through repeated interaction between an agent and an environment. We consider  
 100 the standard formulation in which the agent takes an action  $a_t$  based on a partial observation of  
 101 the current state  $s_t$  the environment is in. The environment transits to the next state  $s_{t+1}$ , and the  
 102 agent is fed with a reward  $r_t$  that acts as the quality assessment of the actions recently taken. This  
 103 repeats until some termination state is reached, the objective of the agent being to determine the  
 104 succession of actions maximizing its cumulative reward over an episode (this is the reference unit  
 105 for agent update, best understood as one instance of the scenario in which it takes actions). Deep  
 106 reinforcement learning (DRL) combines RL and deep neural networks, i.e., collections of connected  
 107 units or artificial neurons, that can be trained to arbitrarily well approximate the mapping function  
 108 between input and output spaces. We consider here fully connected networks in which neurons  
 109 are stacked in layers and information propagates forward from the input layer to the output layer  
 110 via “hidden” layers. Each neuron performs a weighted sum of its inputs to assign significance with  
 111 regard to the task the algorithm is trying to learn, adds a bias to figure out the part of the output  
 112 independent of the input, and feeds an activation function that determines whether and to what  
 113 extent the computed value should affect the outcome.

### 114 B. Proximal policy optimization

115 Proximal policy optimization (PPO) [41] is a model free, on-policy gradient, advantage actor-  
 116 critic reinforcement algorithm. The related key concepts can be summarized as follows:

117 - *model free*: the agent interacts with the environment itself, not with a surrogate model of  
 118 the environment (the corollary here being that it needs no assumptions about the fluid dynamics  
 119 underlying the control problems to be solved).

120 - *policy gradient*: the behavior of the agent is entirely defined by a probability distribution  $\pi(s, a)$   
 121 over actions given states, optimized by gradient ascent. In DRL, the policy is represented by a  
 122 neural network. The free parameters learnt from data are the network weights and biases, with  
 123 respect to which the gradient is computed backwards from the output to the input layer according  
 124 to the chain rule, one layer at the time, using the back-propagation algorithm [62].

125 - *on-policy*: the algorithm improves the policy used to generate the training data (in contrast to  
 126 off-policy methods that also learn from data generated with other policies).

127 - *advantage*: the policy gradient is approximated by that of the policy loss

$$\mathbb{E}_{\tau \sim \pi} \left[ \sum_{t=0}^T \log(\pi(a_t | s_t)) \hat{A}^\pi(s, a) \right], \quad (1)$$

128 where  $\tau = (s_0, a_0, \dots, s_T, a_T)$  is a trajectory of state and actions with horizon  $T$ ,  $A^\pi$  is the advan-  
 129 tage function measuring the gain associated with taking action  $a$  in state  $s$ , compared to taking  
 130 the average over all possible actions, and  $\hat{A}^\pi$  is some biased estimator of the advantage, here its  
 131 normalization to zero mean and unit variance.

132 - *actor-critic*: the learning performance is improved by updating two different networks, a first  
 133 one called actor that controls the actions taken by the agent, and a second one called critic, that  
 134 estimates the advantage as

$$A^\pi(s_t, a_t) = r_t + \gamma V(s_{t+1}) - V(s_t), \quad (2)$$

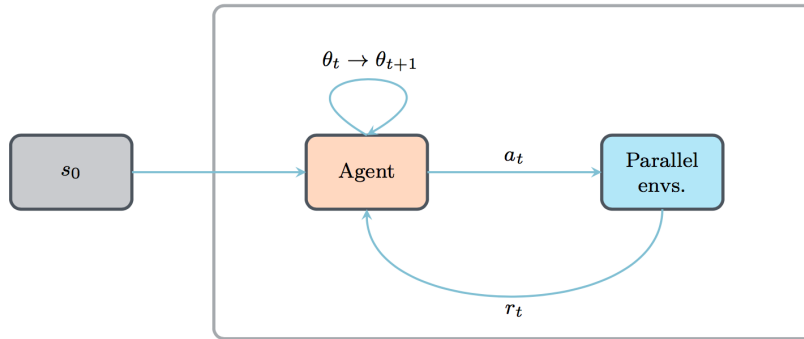


Figure 1. Action loop for single-step PPO. At each episode, the same input state  $s_0$  is provided to the agent, which in turn provides  $n$  actions to  $n$  parallel environments. The latter return  $n$  rewards, that evaluate the quality of each action taken. Once all the rewards are collected, an update of the agent parameters is made using the PPO loss (3). The process is repeated until convergence.

135 where  $V(s)$  is the expected value of the return of the policy in state  $s$  and  $\gamma \in [0, 1]$  is a discount  
 136 factor adjusting the trade-off between immediate and future rewards.

137 PPO uses conservative policy updates to alleviate the issue of performance collapse affecting  
 138 standard policy gradient implementations<sup>1</sup>. We use here PPO-clip<sup>2</sup> to optimize the surrogate loss

$$\mathbb{E}_{(s,a) \sim \pi} \left[ \min \left( \frac{\pi(a|s)}{\pi_{old}(a|s)}, 1 + \epsilon \operatorname{sgn}(\widehat{A}^\pi(s, a)) \right) \widehat{A}^\pi(s, a) \right], \quad (3)$$

139 where  $\epsilon$  is the clipping range defining how far away the new policy is allowed to go from the old.  
 140 The general picture is that a positive (resp. negative) advantage increases (resp. decreases) the  
 141 probability of taking action  $a$  in state  $s$ , but always by a proportion smaller than  $\epsilon$ , otherwise the  
 142 min kicks in (3) and its argument hits a ceiling of  $1 + \epsilon$  (resp. a floor of  $1 - \epsilon$ ). This prevents  
 143 stepping too far away from the current policy, and ensures that the new policy will behave similarly.  
 144 There exist more sophisticated PPO algorithms (e.g., Trust region PPO [63], that determines first  
 145 a maximum step size relevant for exploration, then adaptively adjusts the clipping range to find  
 146 the optimal within this trust region), but standard PPO has simple and effective heuristics. It is  
 147 computationally inexpensive, easy to implement (as it involves only the first-order gradient of the  
 148 policy log probability), and remains regarded as one of the most successful RL algorithms, achieving  
 149 state-of-the-art performance across a wide range of challenging tasks, including flow control [57].

### 150 C. Single-step PPO

151 We now come to single-step PPO (hereafter denoted by PPO-1 to ease the reading), a “de-  
 152 generate” version of PPO introduced in [48] and intended for situations where the optimal policy  
 153 to be learnt by the neural network is state-independent, as is notably the case in open-loop con-  
 154 trol problems (closed-loop control problems conversely require state-dependent policies for which  
 155 standard PPO is best suited). The main difference between standard and single-step PPO can  
 156 be summed up as follows: where standard PPO seeks the optimal set of actions  $a_{opt}$  yielding the  
 157 largest possible reward, single-step PPO seeks the optimal mapping  $f_{\theta_{opt}}$  such that  $a_{opt} = f_{\theta_{opt}}(s_0)$ ,  
 158 where  $\theta$  denotes the network free parameters and  $s_0$  is some input state (usually a vector of zeros)  
 159 consistently fed to the agent for the optimal policy to eventually embody the transformation from  
 160  $s_0$  to  $a_{opt}$ . The agent initially implements a random state-action mapping  $f_{\theta_0}$  from  $s_0$  to an initial  
 161 policy determined by the free parameters initialization  $\theta_0$ , after which it gets only one attempt  
 162 per learning episode at finding the optimal (i.e., it interacts with the environment only once per

<sup>1</sup> Large policy updates can cause the agent to fall off the cliff and to restart from a poorly performing state with a locally bad policy, which is all the more harmful as the step size for policy updating cannot be tuned locally (an above average value can speed up learning in regions of the parameter space where the policy loss is relatively flat, but trigger exploding updates in sharper variation regions).

<sup>2</sup> As opposed to PPO-Penalty, a variant relying on a penalization on the average Kullback–Leibler divergence between the current and new policies, but that tends to perform less well in practice.

163 episode). This is illustrated in figure 1 showing the agent draw a population of actions  $a_t = f_{\theta_t}(s_0)$   
 164 from the current policy, and being returned incentives from the associated rewards to update the  
 165 free parameters for the next population of actions  $a_{t+1} = f_{\theta_{t+1}}(s_0)$  to yield larger rewards.

166 In practice, the agent outputs a policy parameterized by the mean and variance of the probability  
 167 density function of a  $d$ -dimensional multivariate normal distribution, with  $d$  the dimension of the  
 168 action required by the environment. Actions drawn in  $[-1, 1]^d$  are then mapped into relevant  
 169 physical ranges, a step deferred to the environment as being problem-specific. The resolution  
 170 essentially follows the process described in section II B, only the surrogate loss reads

$$\mathbb{E}_{a \sim \pi} \left[ \min \left( \frac{\pi(a)}{\pi_{old}(a)}, 1 + \epsilon \operatorname{sgn}(\widehat{A}^\pi(a)) \right) \widehat{A}^\pi(a) \right], \quad (4)$$

171 and the advantage  $A^\pi$  reduces to the whitened reward  $r_t$ . This is because the trajectory consists  
 172 of a single state-action pair, so the discount factor can be set to  $\gamma = 1$  with no loss of generality. In  
 173 return, the two rightmost terms cancel each other out in (2), meaning that single-step PPO can do  
 174 without the value-function evaluations of the critic network (and is thus not actually actor-critic).

#### 175 D. Computational fluid dynamics environment

176 The CFD resolution framework relies on the in-house, parallel, finite element library Cim-  
 177 LIB\_CFD [64], whose main ingredients are as follows:

178 - the variational multiscale approach (VMS) is used to solve a stabilized weak form of the  
 179 governing equations using linear approximations ( $P_1$  elements) for all variables, which otherwise  
 180 breaks the Babuska–Brezzi condition. The approach relies on an a priori decomposition of the  
 181 solution into coarse and fine scale components [65–67]. Only the large scales are fully represented  
 182 and resolved at the discrete level. The effect of the small scales is encompassed by consistently  
 183 derived source terms proportional to the residual of the resolved scale solution, hence ad-hoc  
 184 stabilization parameters comparable to local coefficients of proportionality.

185 - in laminar regimes, velocity and pressure come as solutions to the Navier–Stokes equations. In  
 186 turbulent regimes, the focus is on phase-averaged velocity and pressure modeled after the unsteady  
 187 Reynolds averaged Navier–Stokes (uRANS) equations. In order to avoid transient negative tur-  
 188 bulent viscosities, negative Spalart–Allmaras [68] is used as turbulence model, whose stabilization  
 189 proceeds from that of the convection-diffusion-reaction equation [69, 70].

190 - the immersed volume method (IVM) is used to immerse and represent all geometries inside a  
 191 unique mesh. The approach combines level-set functions to localize the solid/fluid interface, and  
 192 anisotropic mesh adaptation to refine the mesh interface under the constraint of a fixed, number  
 193 of edges. This ensures that the quality of all actions taken over the course of a PPO optimization  
 194 is equally assessed, even though the interface can depend on the action.

195 Substantial evidence of the flexibility, accuracy and reliability of this numerical framework is  
 196 documented in several papers to which the reader is referred for exhaustive details regarding the  
 197 level-set and mesh adaptation algorithms [71, 72], the VMS formulations, stabilization parameters  
 198 and discretization schemes used in laminar and turbulent regimes [73–76], and the mathematical  
 199 formulation of the IVM in the context of finite element VMS methods [77, 78].

#### 200 E. Numerical implementation

201 In practice, actions are distributed to multiple environments running in parallel, each of which ex-  
 202 ecutes a self-contained MPI-parallel CFD simulation and feeds data to the DRL algorithm (hence,  
 203 two levels of parallelism related to the environment and the computing architecture). Here, all  
 204 CFD simulations are performed on 12 cores of a workstation of Intel Xeon E5-2640 processors.  
 205 The algorithm waits for the simulations running in all parallel environments to be completed, then  
 206 shuffles and splits the rewards data set collected from all environments into several buffers (or  
 207 mini-batches) used sequentially to compute the loss and perform a network update. The process  
 208 repeats for several epochs, i.e., several full passes of the training algorithm over the entire data set,  
 209 which ultimately makes the algorithm slightly off-policy (since the policy network ends up being

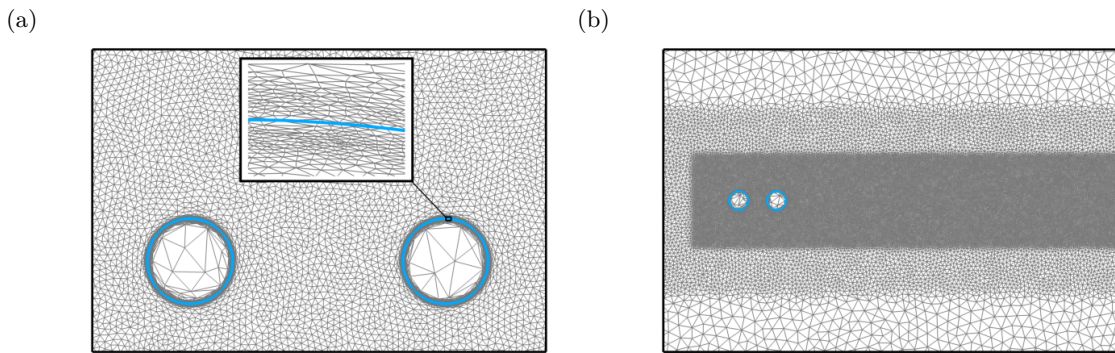


Figure 2. Details of (a) the boundary layer mesh and (b) successive refinement steps of the background mesh for the flow past a tandem arrangement of two circular cylinders. The blue line in (a) indicates the zero iso-contour of the level set function.

Neural network	
2	Nb. hidden layers
4	Nb. neurons/layer
TBS	Nb. epochs
TBS	Nb. environments
TBS	Size of mini-batches
PPO	
$5 \times 10^{-3}$	Learning rate
0.3	Clipping range
1	Discount factor

Table I. Details of the network architecture and PPO hyper parameters. The number of epochs, environments and the size of the mini-batches are provided on a case-by-case basis in sections III and IV.

210 trained on samples generated by older policies, which is customary in standard PPO operation).  
 211 This simple parallelization technique is key to use DRL in the context of CFD applications, as a  
 212 sufficient number of actions drawn from the current policy must be evaluated to accurately esti-  
 213 mate the policy gradient. This comes at the expense of computing the same amount of reward  
 214 evaluations, and yields a substantial computational cost for high-dimensional fluid dynamics prob-  
 215 lems (typically from a few to several hundred CFD simulations for the cases considered herein).  
 216 In the same vein, it should be noted that the common practice in DRL studies to gain insight into  
 217 the performances of the selected algorithm by averaging results over multiple independent training  
 218 runs with different random seeds is not tractable, as it would trigger a prohibitively large compu-  
 219 tational burden. The same random seeds have thus been deliberately used over the whole course  
 220 of study to ensure a minimal level of performance comparison between cases. The remainder of  
 the practical implementation details are as follows:

221  
 222 - the environment consists of CFD simulations of two-dimensional (2-D) flows described in a  
 223 Cartesian coordinate system with drag positive in the  $+x$  direction. All equations are discretized  
 224 on rectangular grids whose side lengths documented in the coming sections have been checked to  
 225 be large enough not to have a discernible influence on the results (with the exception of the square  
 226 cylinder flow in section IV A 3 and the fluidic pinball in section IV B, for which we use respectively  
 227 the values recommended in [79] and the same values as in [34]). Open flow conditions are used, that  
 228 consist of a uniform inflow in the  $x$  direction, together with symmetric lateral, advective outflow  
 229 and no-slip interface conditions. In turbulent regime, the ambient value of the Spalart–Allmaras  
 230 variable is three times the molecular viscosity, as recommended to lead to immediate transition.  
 231 Typical adapted meshes of the interface and wake regions are shown in figure 2, the latter also  
 being accurately captured via successive refinement of the background elements.

232  
 233 - the instant reward is (up to a plus/minus sign) either the time-averaged or the root mean  
 234 square (rms) value of the force coefficient (drag or lift per unit span length), to consider either  
 235 the mean or fluctuating force acting on the immersed body. Instantaneous values are computed

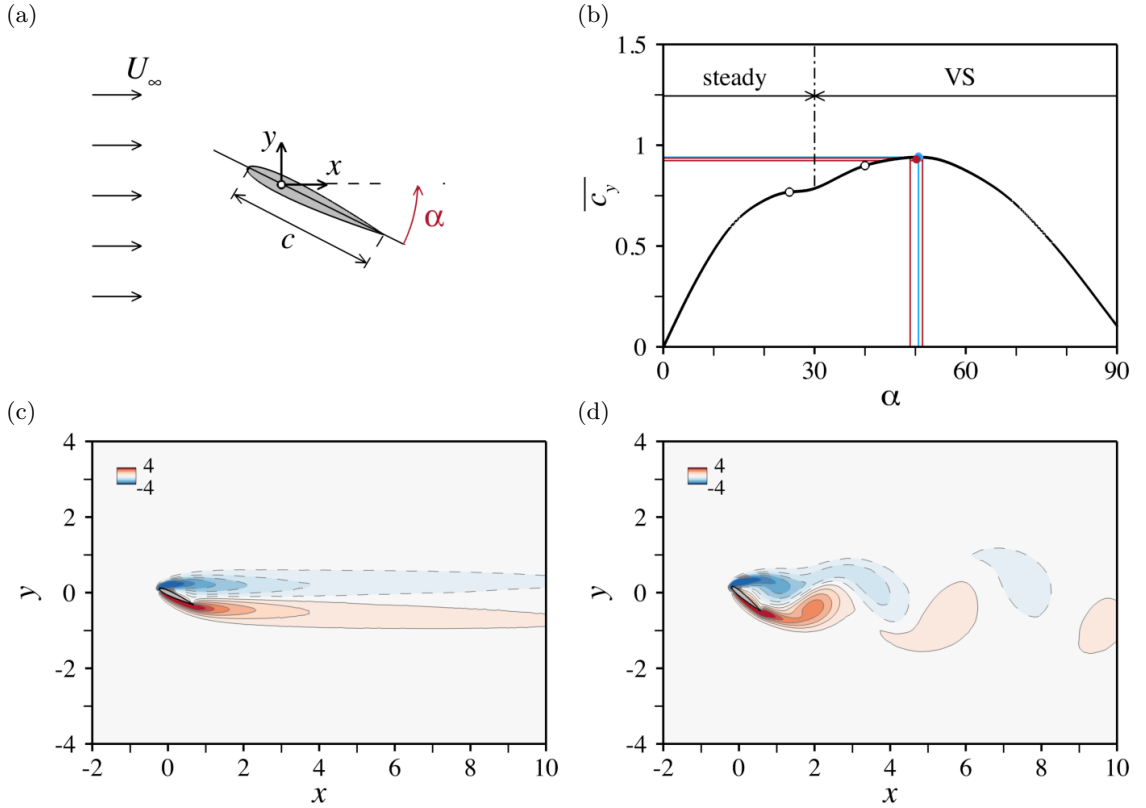


Figure 3. Flow past a NACA 0012 - (a) Schematic diagram of the configuration. (b) Mean lift against the angle of attack computed by DNS at  $Re = 100$ . The VS label indicates the angles for which the flow exhibits unsteadiness in the form of periodic vortex formation and shedding. The blue lines and symbols mark the optimal. The red symbol is the average over the 5 latest single-step PPO episodes, and the red lines delimit the corresponding variance intervals. (c-d) Instantaneous vorticity fields computed at  $Re = 100$ , for values marked by the circle symbols in (b), namely (c)  $\alpha = 25$  and (d)  $\alpha = 40$ .

236 with a variational approach featuring only volume integral terms, reportedly less sensitive to the  
 237 approximation of the body interface than their surface counterparts [80, 81]. Time averages are  
 238 performed over an interval  $[t_i; t_f]$  with edges large enough to dismiss the initial transient and achieve  
 239 convergence to statistical equilibrium. Moving average rewards and actions are also computed as  
 the sliding average over the 50 latest values (or the whole sample if it has insufficient size).

240  
 241 - the agent is a fully connected network with 2 hidden layers, each of which holds 4 neurons with  
 242 hyperbolic tangent activation functions. We use the default online PPO implementation of Stable  
 243 Baselines, a toolset of reinforcement learning algorithms dedicated to the research community  
 244 and industry [82], for which a custom OpenAI environment has been designed with the Gym  
 245 library [83]. Unlike other RL algorithms, PPO does not generally require significant tuning of the  
 246 hyper parameters (i.e., parameters that are not estimated from data). Nonetheless, all values used  
 247 in this study are documented in table I to ease reproducibility, including the learning rate (the size  
 248 of the step taken in the gradient direction for policy update), the PPO clipping range (set to the  
 249 upper edge of the recommended range) and the discount factor (set to the default PPO-1 value).

250

### III. APPLICATION TO FLOW OPTIMIZATION

251

#### A. Flow past a NACA 0012 airfoil

252 We consider first a NACA 0012 airfoil placed at incidence in a uniform stream, as depicted in  
 253 figure 3(a). The origin of the coordinate system is at the airfoil pivot-point, set at quarter chord  
 254 length from the leading edge. A laminar, time-dependent case at Reynolds number  $Re = U_\infty c / \nu =$



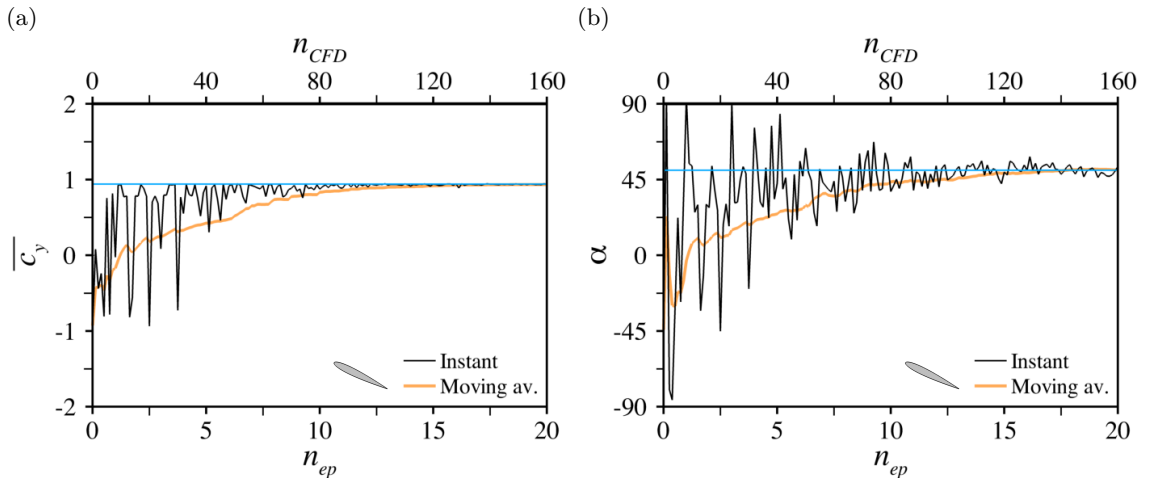


Figure 4. Flow past a NACA 0012 at  $Re = 100$  - (a) Evolution per episode for the instant (black line) and moving average (over episodes, orange line) values of the mean lift (over time). The corresponding number of CFD simulations (obtained multiplying by the number of environments) is displayed on the secondary horizontal axis. (b) Same as (a) for the angle of attack. The blue lines mark the DNS optimal.

100 is modeled after the Navier–Stokes equations, where  $U_\infty$  is the inflow velocity,  $c$  the straight chord distance and  $\nu$  the kinematic viscosity. The objective is to maximize the mean lift  $\bar{c}_y$ , for which the sole control parameter is the angle of attack  $\alpha$  measuring the incidence relative to the chord (in degrees and with the convention that  $\alpha > 0$  for the airfoil to generate positive lift. Also, we keep in mind that  $\alpha$  is rather a state parameter than an adjustable control parameter in practical situations, but the methodology carries over to related optimization problems such as the design of multi-element high-lift systems). This is a problem simple enough to allow direct comparisons between PPO-1 and DNS (actually VMS, but the difference is clear from context), all the more so as lift varies smoothly with the incidence. This is evidenced in figure 3(b) showing reference data obtained from 15 DNS runs computing the mean lift to an accuracy of 3% with the simulation parameters documented in table II. The distribution changes slope near  $\alpha \sim 30^\circ$  (because the system bifurcates from a steady to a time-periodic vortex-shedding regime; see figure 3(c-d) showing instantaneous vorticity fields computed on either side of the threshold) but otherwise exhibits a well-defined, smooth maximum at  $\alpha^* = 50.6$ , associated with  $\bar{c}_y^* = 0.94$ .

For each PPO-1 learning episode, the network outputs a single value  $\xi$  in  $[-1; 1]$  mapped into

$$\alpha = \xi \alpha_{\max}, \quad (5)$$

for the angle of attack to vary in  $[-\alpha_{\max}; \alpha_{\max}]$  with  $\alpha_{\max} = 90^\circ$ . The reward  $r = \bar{c}_y$  is then computed using the same simulation parameters, after which the network is updated for 32 epochs using 8 environments and 4 steps mini-batches. 20 episodes have been run for this case, which represents 160 simulations, each of which lasts  $\sim 25$ mn using 12 cores,<sup>3</sup> hence  $\sim 65$ h of total CPU cost (equivalently,  $\sim 8$ h of resolution time). We show in figure 4(a) the evolution of the reward collected over the course of the optimization. The moving average increases almost monotonically and reaches a plateau after about 15 episodes, and the optimal lift computed as the average over the 5 latest episodes is  $\bar{c}_y^* = 0.93 \pm 0.01$  (the variations are computed from the rms of the moving average over the same interval, which is a simple yet robust criterion to assess qualitatively convergence a posteriori). The associated angle  $\alpha^* = 50.2^\circ \pm 1.2^\circ$  varies by a larger factor, which is because lift is relatively insensitive to the exact incidence in the vicinity of the optimal. This is perfectly in line with the DNS, as illustrated by the red lines in figure 3(b) showing the limits of the so-computed variance intervals. Nonetheless, PPO-1 turns to be rather inefficient at finding the optimal, because it must span continuous ranges of angles while the one-dimensionality of the control space and the smoothness of the optimal allow DNS to test only a few discrete values (hence it can converge within  $\sim 1$ h using the same level of CFD parallelization).

<sup>3</sup> This is the time needed to compute periodic vortex shedding solutions. It takes less than 10mn to march the solution to steady state, but this barely affects the total CPU cost, as the time needed to complete an episode is that of completing its longest simulation (so only the cost of those episodes exclusively computing steady state solutions is reduced by a few minutes).

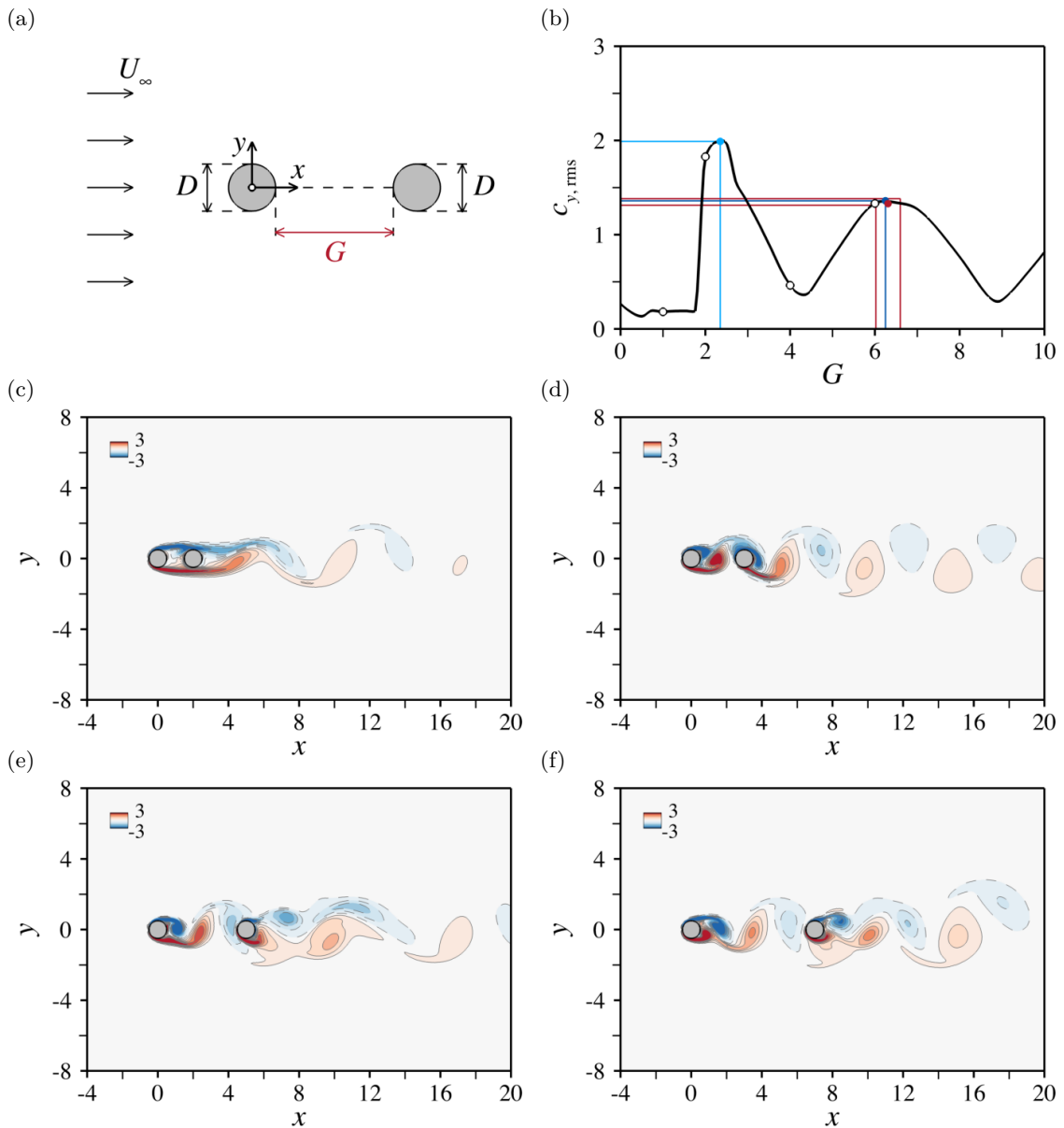


Figure 5. Flow past the tandem arrangement of two circular cylinders - (a) Schematic diagram of the configuration. (b) Fluctuating (rms) lift against the gap spacing computed by DNS at  $Re = 300$ . The red symbol is the average over the 5 latest single-step PPO episodes, and the red lines delimit the corresponding variance intervals. (c-f) Instantaneous vorticity fields computed at  $Re = 300$ , for values marked by the circle symbols in (b), namely (c)  $G = 1$ , (d)  $G = 2$ , (d)  $G = 4$ , and (e)  $G = 6$ .

286

### B. Flow past an arrangement of two side-by-side circular cylinders

287 We examine now the side-by-side tandem arrangement of two identical circular cylinders in a  
 288 uniform stream, whose configuration is sketched in figure 5(a). The origin of the coordinate  
 289 system is at the center of the main cylinder, where we refer to the upstream and downstream cylinders as  
 290 “main” and “surrounding”, respectively. A laminar, time-dependent case at  $Re = U_\infty D/\nu = 300$   
 291 is modeled after the Navier–Stokes equations, where  $D$  is the diameter of either cylinder. The  
 292 objective is to maximize the rms lift  $c_{y,rms}$  of the two-cylinder system (for instance, to increase  
 293 the amount of energy available for harnessing from fluid-structure interactions) for which the sole  
 294 control parameter is the gap spacing  $G$ , i.e., the side-to-side distance between the two cylinders.  
 295 On paper, this is another problem simple enough to allow direct comparisons between PPO-1 and

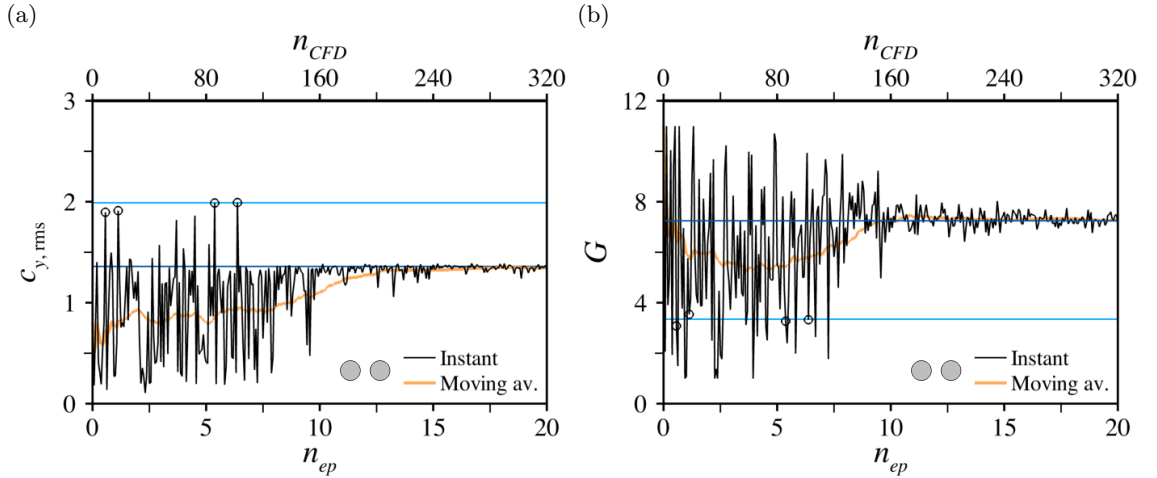


Figure 6. Flow past the tandem arrangement of two circular cylinders at  $Re = 300$  - (a) Evolution per episode for the instant (black line) and moving average (over episodes, orange line) values of the rms lift. (b) Same as (a) for the gap spacing. The light (resp. dark) blue lines mark the DNS global (resp. local) maximum. The circles are high reward parameters close to the DNS global maximum.

296 DNS. In practice, the results are not so unequivocal, as evidenced in figure 5(b) showing reference  
 297 data obtained from 30 DNS runs computing the rms lift to an accuracy of 5% with the simulation  
 298 parameters documented in table II. A steep global maximum lies at  $G^* = 2.35$ , associated with  
 299  $c_{y,rms}^* = 1.99$ , but there is a smoother local maximum at  $G^{**} = 6.25$ , associated with  $c_{y,rms}^{**} = 1.36$ ,  
 300 which reflects the high sensitivity of the pattern of flow unsteadiness to the center distance. Namely  
 301 without going into too much detail (as this has been extensively discussed in the literature [84–87]),  
 302 the instantaneous vorticity field computed for  $G = 1$  in figure 5(c) shows that the gap flow between  
 303 the two cylinders is initially steady, while the shear layers separating from the main cylinder engulf  
 304 those of the surrounding cylinder and trigger vortex shedding in the far wake. For  $G = 2$  (close to  
 305 the global maximum), the gap flow is unsteady, but the gap vortices are not fully developed by the  
 306 time they impinge on the surrounding cylinder, hence a single vortex street in the far wake; see  
 307 figure 5(d). For  $G = 4$ , one pair of gap vortices fully develops, then impinges on the surrounding  
 308 cylinder, which triggers a complex interaction in the near wake before a vortex street eventually  
 309 forms further downstream; see figure 5(e). Finally for  $G = 6$  (close to the local maximum) the  
 310 wake of the surrounding cylinder is unsteady again, and both cylinders shed synchronized vortices  
 311 close to anti-phase; see figure 5(f).

312 For each PPO-1 learning episode, the network outputs a single value  $\xi$  in  $[-1; 1]$  mapped into

$$G = \frac{1 + \xi}{2} G_{\max}, \quad (6)$$

313 for the gap to vary in  $[0; G_{\max}]$  with  $G_{\max} = 10$ . This enables contact between the two cylinders and  
 314 keeps the computational cost affordable, as pushing the surrounding cylinder further downstream  
 315 would require extending the computational domain and increasing the numbers of grid points  
 316 accordingly (all the more so as we do not anticipate such large distances to be relevant from the  
 317 standpoint of optimization because the interaction between both cylinders will weaken increasingly  
 318 at some point, although it can take up to several tens of diameters to do so). The reward  $r = c_{y,rms}$   
 319 is then computed using the same simulation parameters, after which the network is updated for 32  
 320 epochs using 16 environments and 4 steps mini-batches. Another 20 episodes have been run for this  
 321 case. This represents 320 simulations, each of which lasts  $\sim 60$ mn on 12 cores (much longer than in  
 322 the NACA case due to the increased simulation time), hence  $\sim 320$ h of total CPU cost (equivalently,  
 323  $\sim 20$ h of resolution time), still much more than by DNS because DRL keeps spanning continuous  
 324 ranges of distances while DNS can settle for only a few discrete values despite the sharpness of the  
 325 global maximum (hence it can converge within  $\sim 3$ h using the same level of CFD parallelization).  
 326 Figure 6(a) shows a plateau in the moving average reward after about 15 episodes. The optimal  
 327 lift computed as the average over the 5 latest episodes is  $c_{y,rms}^* = 1.34 \pm 0.02$ , associated with  
 328  $G^* = 6.31 \pm 0.04$ , meaning that the agent misses the global maximum, but converges to a value  
 329 close to the local maximum; see the red lines in figure 5(b) indicating the limits of the computed

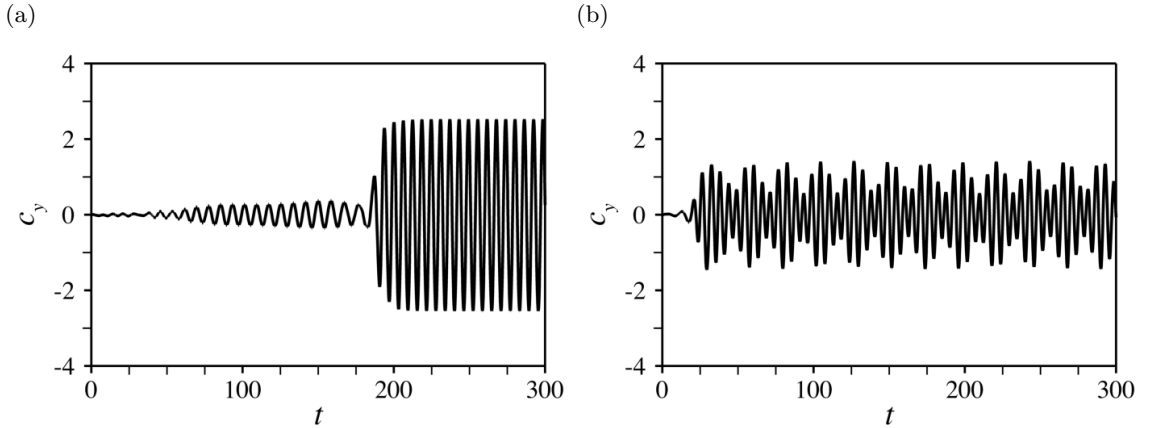


Figure 7. Flow past the tandem arrangement of two circular cylinders at  $Re = 300$  - Time history of lift computed by DNS for gap spacings (a)  $G = 2$  (close to the DNS global maximum) and (b)  $G = 8$ .



	$\bar{c}_y$		$\alpha$		$c_{y,rms}$	$G$		
	0.93	50.2°	PPO-1		1.34	6.31	PPO-1	
	0.94	50.6°	DNS		1.99	2.35	DNS	Optimal
					1.36	6.25	DNS	
CFD								
		100				300		Reynolds number
		0.125				»		Time-step
		[50; 150]				[200; 300]		Averaging time span
		$[-15; 40] \times [-15; 15]$				»		Mesh dimensions
		115000				125000		Nb. mesh elements
		0.001				»		Interface $\perp$ mesh size
		12				»		Nb. Cores
PPO-1								
		20				»		Nb. DRL episodes
		8				16		Nb. Environments
		32				»		Nb. Epochs
		4				»		Size of mini-batches
		60h				320h		CPU time
		7.5h				20h		Resolution time

Table II. Simulation parameters and convergence data for the flow past a NACA 0012 at  $Re = 100$  and the flow past the tandem arrangement of two circular cylinders at  $Re = 300$ . NACA 0012: the interface mesh size yields  $\sim 20$  grid points in the boundary-layer at mid-chord, under zero incidence, and the averaging time-span represents  $\sim 15 - 20$  shedding cycles, depending on the incidence. Tandem arrangement of two circular cylinders: the interface mesh size yields  $\sim 20$  grid points in the boundary-layer of the main cylinder, just prior to separation, and the averaging time-span represents  $\sim 20$  shedding cycles.

330 variance intervals.

331 This half-failure can be explained by the steepness of the reward gradients with respect to the  
 332 control variable in the vicinity of the global maximum. This is due to the existence of a secondary  
 333 instability mechanism at play in a narrow range of center distances, as illustrated in figure 7(a)  
 334 showing that for  $G \sim 2$ , the flow settles to a first time-periodic solution, then bifurcates to a  
 335 second time-periodic solution associated with increased lift oscillations (hence the large values of  
 336  $t_i$  used for this case). Actually, DRL does identify high reward positions close to  $G = 2$  (circle  
 337 symbols in figure 6), whose value  $c_{y,rms} \sim 2$  is consistent with the global maximum, but there are  
 338 very few times where the global maximum is met during the exploration phase (compared to its  
 339 local counterpart, again because of the topology of the reward function). Because PPO voluntarily  
 340 dismisses large policy updates to avoid performance collapse, the clipped policy updates only lead  
 341 to limited exploration and trap the optimization process into a local maximum. Low to moderate

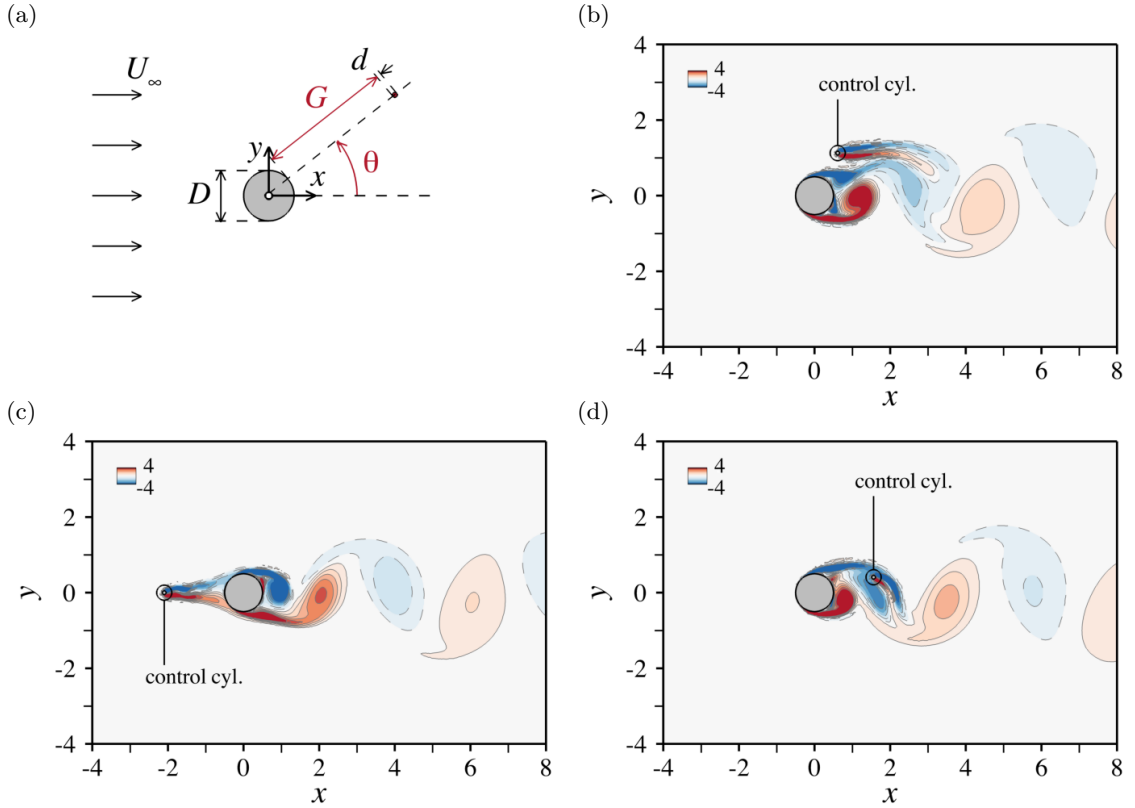


Figure 8. Open-loop control of the circular cylinder flow by a small control cylinder of diameter  $d = 0.1$  - (a) Schematic diagram of the configuration. (b-d) Iso-contours of the vorticity field computed at  $Re = 3900$  for representative positions  $(x_c, y_c)$  of the control cylinder, namely (b)  $(0.61, 1.13)$ , (c)  $(-2.10, 0.00)$  and (d)  $(1.56, 0.41)$ .

342 Reynolds numbers are likely required for such instability cascade scenario to occur, so such results  
 343 do not cast doubt on the applicability of single-step PPO to practically meaningful high Reynolds  
 344 flows. They do stress, however, that the method can benefit from carefully tuning the trade-off  
 345 between exploration and exploitation, which will be addressed in future work.

#### 346 IV. APPLICATION TO OPEN-LOOP FLOW CONTROL

##### 347 A. Optimal cylinder drag reduction using a smaller control cylinder

348 The relevance of single-step PPO is now showcased by tackling various open-loop control prob-  
 349 lems. The first one is that of a cylinder in a uniform stream, controlled open-loop by a much  
 350 smaller circular cylinder. Figure 8(a) presents a sketch of the configuration pertaining to a cir-  
 351 cular geometry of the main cylinder, where we refer to the large and small cylinders as “main”  
 352 and “control”, respectively, but section IV A 4 also considers a square geometry. The origin of the  
 353 coordinate system is at the center of the main cylinder. The objective is to minimize the mean drag  
 354  $\overline{c_x}$  of the two-cylinder system, which requires reducing the drag of the main cylinder sufficiently  
 355 to compensate for the fact that the control cylinder itself is a source of drag. Several laminar and  
 356 turbulent Reynolds numbers  $Re = U_\infty D / \nu$  are considered, where  $D$  is the diameter of the main  
 357 cylinder. The diameter of the control cylinder is set to  $d = 0.1$ , therefore the sole control parameter  
 358 is the 2-D position of the control cylinder center, measured by the gap distance  $G$  between the  
 359 two cylinders and the azimuthal position  $\theta$  with respect to the rear stagnation point. This may  
 360 not seem overly complicated on paper, but the parameter space is actually large enough to dismiss  
 361 mapping the best positions for placement of the control cylinder by DNS, as tens of thousands  
 362 of runs are required to cover merely a few diameters around the main cylinder. In the following,

single-step PPO is thus compared to theoretical predictions obtained by the adjoint method. The latter has proven fruitful to gain insight into the most efficient region from the linear sensitivity of the uncontrolled flow (i.e., the flow past the main cylinder), without ever calculating the controlled states, using instead a simple model of the force exerted by the control cylinder on the flow. We shall not go into the technicalities of how to derive the related adjoint equations, as the line of thought here is to take the output sensitivity as a given to assess relevance of PPO-1. Suffice it to say here that we rely on various levels of adjoint modeling whose key assumptions are reviewed in appendix A. The reader interested in more details is directed to the original literature on this topic [25, 88, 89], where in-depth technical and mathematical information, together with extensive discussions regarding the validity of the approximations are available. From the numerical standpoint, all calculations are performed with the mixed finite elements adjoint solver presented and validated in [25].

On the CFD side, one of the challenges lies in the fact that the control cylinder acts as a small local disturbance redistributing the vorticity in the separated shear layers; see figures 8(b-d) showing instantaneous vorticity fields computed for representative positions of the control cylinder. Accurate numerical methods are thus mandatory to capture the small drag variations induced by the control. Several values of the Reynolds number are investigated : a laminar, steady case at  $Re = 40$ , for which the flow remains steady-state regardless of the position of the control cylinder, a laminar, time-dependent case at  $Re = 100$ , for which vortex shedding consistently develops from the main cylinder but the flow past the control cylinder remains steady, and two turbulent cases at  $Re = 3900$  and at  $Re = 22000$  (hence modeled after the uRANS equations with negative Spalart–Allmaras as turbulence model), for which vortex shedding develops from both cylinders. This is because the Reynolds number in the wake of the control cylinder must be scaled by the ratio of the cylinder diameters, which yields values below (resp. above) the instability threshold at  $Re = 100$  (resp.  $Re = 3900$  and  $Re = 22000$ ).

For each PPO-1 episode, the network outputs two values  $\xi_{1,2}$  in  $[-1; 1]^2$  mapped into

$$G = \frac{1 + \xi_1}{2} G_{\max}, \quad \theta = \frac{1 + \xi_2}{2} \theta_{\max}, \quad (7)$$

for the gap to vary in  $[0; G_{\max}]$  with  $G_{\max} = 3$ , and the azimuthal position to vary in  $[0; \theta_{\max}]$  with  $\theta_{\max} = 180^\circ$ . This enables contact between the two cylinders, and allows taking advantage of the problem symmetry, as it amounts to moving the control cylinder in the upper half of a torus bounded by the surface of the main cylinder and the user-defined exterior radius  $G_{\max}$ . In the following, the center position is conveniently presented in terms of the Cartesian coordinates  $x_c = \rho \cos \theta$  and  $y_c = \rho \sin \theta$ , where we note  $\rho = G + (1 + d)/2$ . Since the aim is to minimize drag, the reward  $r = -\overline{D}$  is then computed using the simulation parameters documented in table III, after which the network is updated for 32 epochs using 8 environments and 2 steps mini-batches (note the zero averaging span in table III for  $Re = 40$ , as this is a steady case for which the steady asymptotic value of total drag can be evaluated at the final time  $t_f$ , provided it is large enough for the solution to relax to steady-state).

#### 1. Laminar steady regime and circular geometry at $Re=40$

For this first case, 100 episodes have been run, which represents 800 simulations, each of which lasts  $\sim 35$ mn on 12 cores, hence  $\sim 480$ h of total CPU cost (equivalently,  $\sim 60$ h of resolution time). The moving average value of drag reaches a plateau after about 60 episodes in figure 9(a), with the optimal value  $\overline{c_x}^* = 1.53 \pm 0.01$  computed as the average over the 5 latest episodes representing a reduction by roughly 2% with respect to the uncontrolled value 1.56 (in good agreement with the reference 1.54 from the literature [90, 91]). Meanwhile, the instant value of drag actually keeps oscillating over the next 40 episodes with small but finite amplitude, which is further evidenced in figure 9(b-c) showing the instant and moving average center positions of the control cylinder. On the one hand,  $y_c^*$  quickly settles to zero, i.e., the control cylinder converges to the horizontal centerline. On the other hand,  $x_c$  keeps exchanging positions between two regions distributed almost symmetrically on either side of the main cylinder, an upstream region associated with  $\overline{c_x} \sim 1.51$  and a slightly less efficient downstream region associated with  $\overline{c_x} \sim 1.54$ , which suggests that the drag functional has global and local minima located in valleys of comparable depth. Confirmation comes from the theoretical drag variations computed (in steady mode) from the

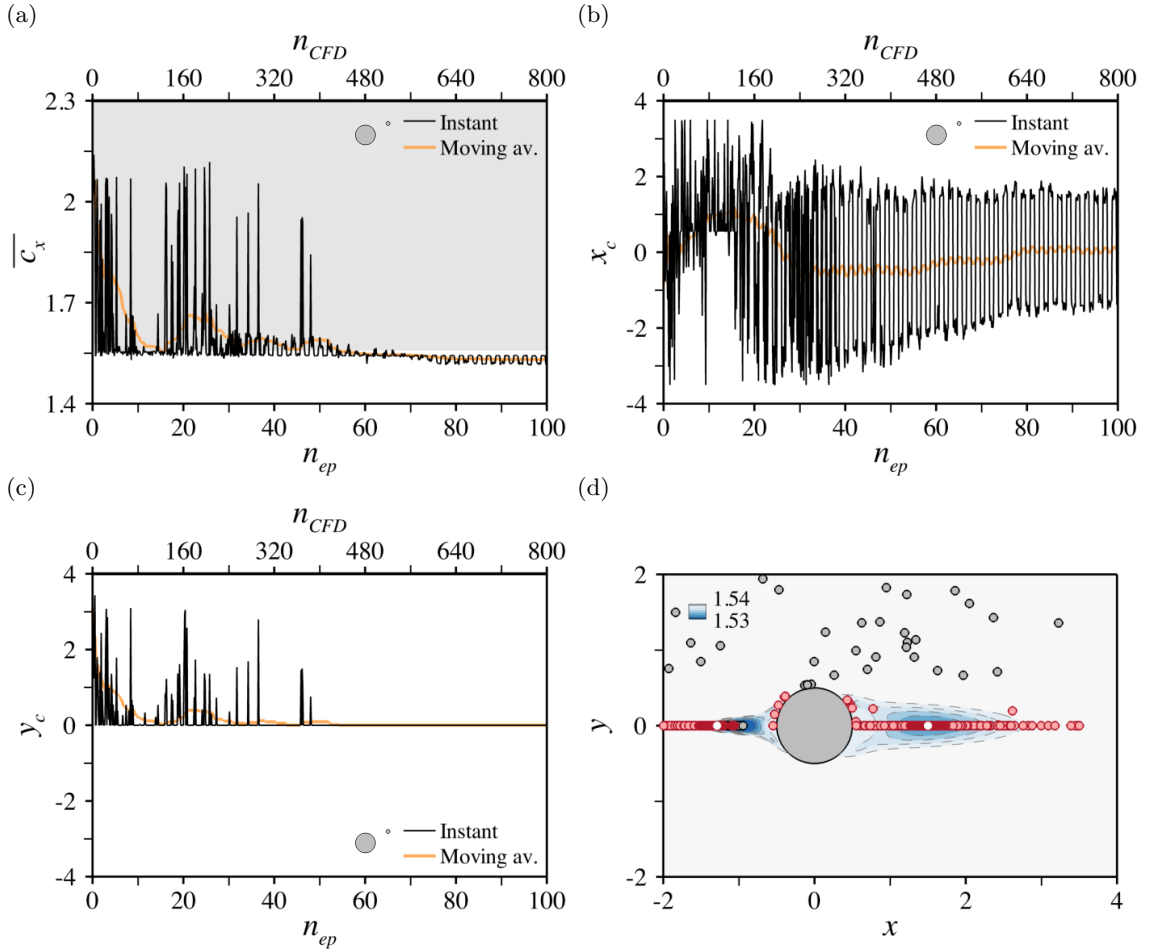


Figure 9. Open-loop control of the circular cylinder flow by a small control cylinder of diameter  $d = 0.1$  at  $Re = 40$  - (a) Evolution per episode for the instant (black line) and moving average (over episodes, orange line) values of the mean drag (over time). The uncontrolled drag is at the bottom of the grey shaded area. (b-c) Same as (a) for the  $x_c$  and (c)  $y_c$  positions of the control cylinder center. (d) Theoretical mean drag variation computed by a steady adjoint method modeling the presence of the control cylinder by a pointwise reacting force localized at the same location where the control cylinder is placed (only the negative iso-contours are reported for clarity). The grey circles are the positions investigated by the DRL. The light red circles are high reward positions spanned over the course of optimization. The dark red circles are those high reward positions spanned over the last 5 episodes. The white circles are the median values reported in the summarizing table III.

415 baseline adjoint method described in appendix A 1, whose negative iso-values (associated to drag  
 416 reduction) are mapped in figure 9(d). The latter unveil two regions nestled against either side of  
 417 the main cylinder and achieving similar drag reduction by  $\sim 2\%$ , a first one extending upstream  
 418 over approximately 1 diameter, and a second one, slightly less efficient and extending downstream  
 419 and along the outer boundary of the recirculation over 3 diameters. DRL manages to find high-  
 420 reward positions in both, which is best seen from the various symbols in figure 9(d) showing the  
 421 complete set of PPO-1 positions investigated over the course of optimization (grey circles) together  
 422 with those positions achieving optimal drag reduction within 5% (light red circles), including  
 423 a few non-centerline positions along the edge of both drag reduction regions. Nonetheless, the  
 424 algorithm ultimately converges to almost symmetrical core positions, as evidenced by the dark red  
 425 circles in figure 9(d) showing the positions spanned over the 5 latest episodes. Despite limited  
 426 discrepancies regarding the exact position of the upstream region (slightly shifted upstream in the  
 427 present approach), this is consistent with the adjoint-based results and clearly assesses the ability  
 428 of single-step PPO to identify both regions of interest and to accurately predict the drag reduction  
 429 achieved in these regions.

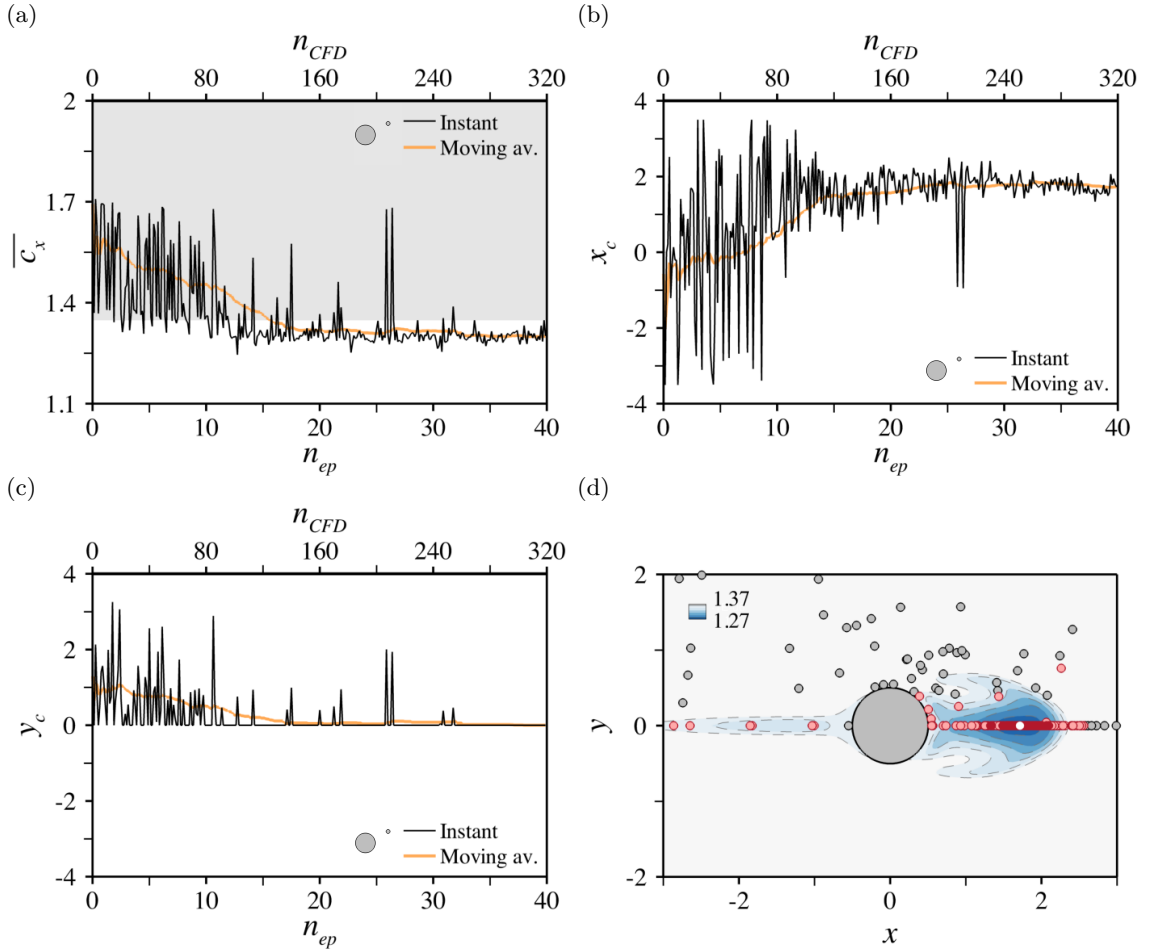


Figure 10. Open-loop control of the circular cylinder flow by a small control cylinder of diameter  $d = 0.1$  at  $Re = 100$  - Same as figure 9, only the theoretical variations in (d) have been computed by the time-varying adjoint method presented in [25].

430

## 2. Laminar time-dependent regime and circular geometry at $Re=100$

431 For this case, 40 episodes have been run, which represents 320 simulations, each of which lasts  
 432  $\sim 1$ h on 12 cores, hence  $\sim 320$ h of total CPU cost (equivalently,  $\sim 40$ h of resolution time). The  
 433 moving average reward plateaus after about 25 episodes in figure 10(a), with the optimal drag  $\bar{c}_x^* =$   
 434  $1.30 \pm 0.01$  computed as the average over the 5 latest episodes representing a reduction by roughly  
 435 5% with respect to the uncontrolled value 1.37 (close to the reference 1.35 from the literature  
 436 [91]). Unlike the previous steady case at  $Re = 40$ , the center position of the control cylinder  
 437 exhibits a similarly converging behavior in figure 10(b-c) with  $x_c^* = -1.76 \pm 0.03$  and  $y_c^* = 0$ ,  
 438 which suggests that the drag functional now has a well-defined global minimum. Confirmation  
 439 comes from the theoretical drag variations computed (in unsteady mode) from the baseline adjoint  
 440 method, whose negative iso-values mapped in figure 10(d) are reproduced from [92]. The latter  
 441 unveil again two regions nestled against either side of the main cylinder, a first one extending  
 442 upstream over approximately 2 diameter (more than at  $Re = 40$ ), and a second one extending  
 443 downstream and along the outer boundary of the mean recirculation over 2 diameters (less than  
 444 at  $Re = 40$ ). Drag is reduced by roughly 2% upstream, but almost 8% downstream, meaning  
 445 that the drag functional has global and local minima in valleys of different depth, in line with the  
 446 DRL results. Again, DRL finds high-reward positions in both regions, as evidenced in figure 10(d)  
 447 by the complete set of PPO-1 positions investigated over the course of optimization (small grey  
 448 circles) and the positions achieving optimal drag reduction within 5% (light red circles), including  
 449 a few centerline upstream positions. The algorithm however quickly settles for the most efficient  
 450 downstream region, as the positions spanned over the 5 latest episodes (dark red circles) all lie in



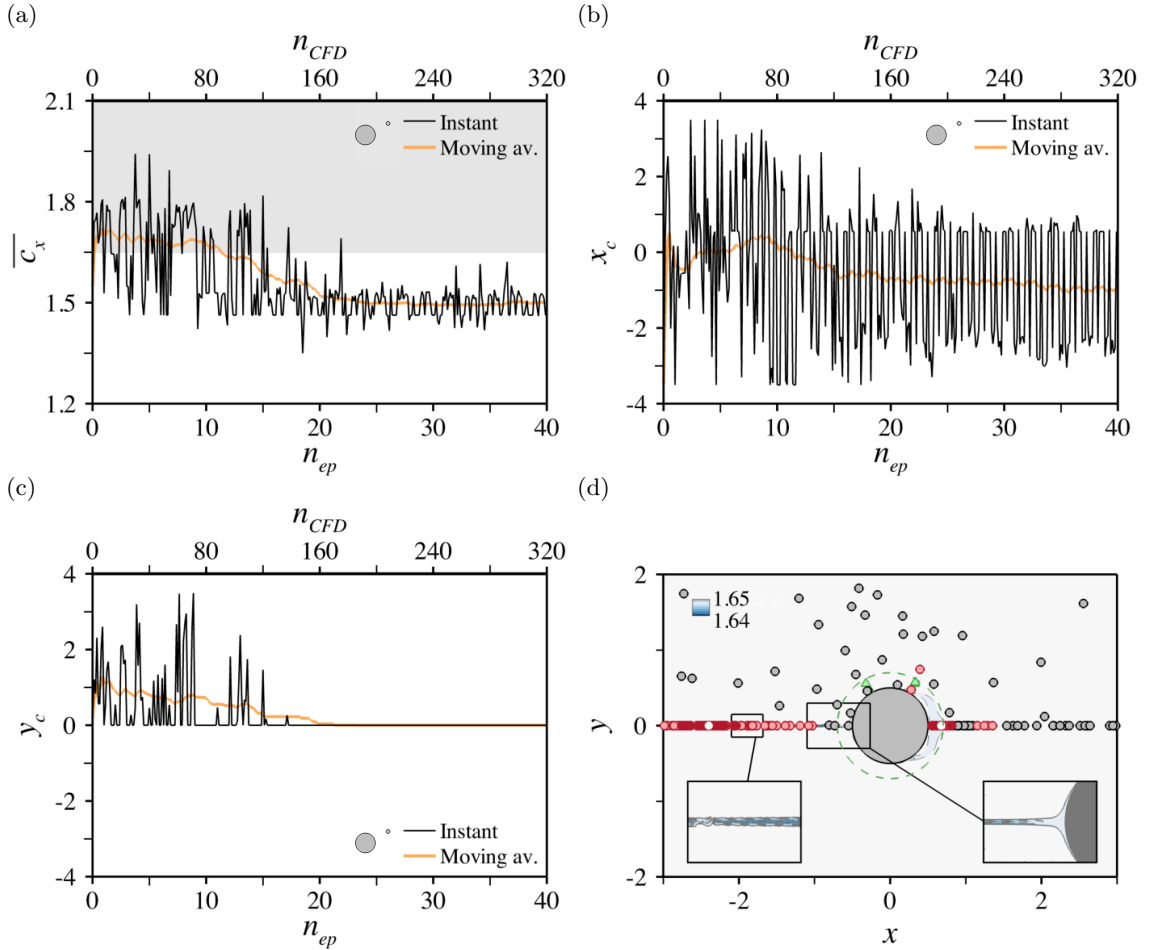


Figure 11. Open-loop control of the circular cylinder flow by a small control cylinder of diameter  $d = 0.1$  at  $Re = 3900$  - Same as figure 9, only the theoretical variations in (d) have been computed by the (steady) simplified mean-flow adjoint method presented in [25]. The green dashed circle in (d) indicates the range of center positions spanned experimentally in [93], with the green triangles marking the sets of positions found to optimally reduce the drag of the main cylinder only.

451 the core of the mean recirculation region, in striking agreement with the adjoint-based results.

452

### 3. Turbulent regime and circular geometry at $Re=3900$

453 Another 40 episodes have been run for this case, which represents 320 simulations, each of which  
 454 lasts  $\sim 2\text{h}30$  on 12 cores (much longer than at  $Re = 100$  due to the halved time step), hence  
 455  $\sim 800\text{h}$  of total CPU cost (equivalently,  $\sim 100\text{h}$  of resolution time). After about 20 episodes, the  
 456 moving average reward in figure 11(a) converges to  $\bar{c}_x^* = 1.50 \pm 0.01$ , which represents a reduction  
 457 of drag by 9% with respect to the uncontrolled value 1.65 (in good agreement with reference 2-D  
 458 RANS data from the literature [94]). The center position of the control cylinder however keeps  
 459 oscillating over the next 15 episodes in figure 11(b-c), as  $y_c^*$  goes to zero but  $x_c$  exchanges positions  
 460 between two regions located on either side of the main cylinder, an upstream region associated  
 461 with  $\bar{c}_x \sim 1.52$  and a downstream region associated with  $\bar{c}_x \sim 1.46$ . This suggests that the drag  
 462 functional has global and local minima in valleys of comparable depth, which is reminiscent of the  
 463 steady case at  $Re = 40$ , only the deepest valley is now downstream, not upstream. Interestingly,  
 464 Ref. [93] determines experimentally different optimal positions  $(G, \theta) = (0.14 - 0.16, 60^\circ)$  and  
 465  $(0.06 - 0.14, 115^\circ)$ , shown as the green triangles in figure 11(d). Additional DNS runs have thus  
 466 been carried out to confirm sub-optimality for our case, although the algorithm does identify  
 467 a couple of high-reward positions in the vicinity of the downstream experimental region. This

468 probably stems from the noticeable differences between both studies, as the Reynolds number  
 469 in [93] is larger by one order of magnitude ( $Re = 65000$ ), the control cylinder is almost twice as  
 470 small ( $d = 0.06$ ), and the experiments focus on the drag of the main cylinder (not the total drag)  
 471 while spanning a much smaller range of center positions (indicated by the green dashed circle in  
 472 figure 11(d)).

473 The DRL results are conversely qualitatively in line with the negative iso-values of the adjoint-  
 474 based drag variations shown in figure 11(d). Those indicate that drag is reduced in two distinct  
 475 regions nestled against either side of the main cylinder, a first narrow one extending upstream  
 476 along the centerline over approximately 2 diameters, and a second one extending downstream over  
 477 a half-diameter and in the vicinity of the mean separation points. Nonetheless, the agreement is not  
 478 quantitative, as the theoretical variations are by a mere 1% upstream (and even lower downstream).  
 479 This is most likely because all theoretical variations have been modeled after a simplified adjoint  
 480 method intended to guide near-optimal design with marginal computational effort (as it requires  
 481 knowledge of the sole mean uncontrolled solution, as explained in appendix A 2), that ends up  
 482 miscalculating the effect of the control cylinder because of an insufficient level of sophistication. On  
 483 the one hand, the marginal size of the downstream region (as well as the marginal drag reduction  
 484 predicted in this region) is ascribed to the fact that the approach has been shown to possibly  
 485 miss out on sensitivity regions involving strong interactions of the mean and fluctuating solution  
 486 components via the formation of Reynolds stresses [92]: the mean recirculation is one such region  
 487 where reducing the drag of the main cylinder, even by a small amount, suffices to reduce the total  
 488 drag because the  $x$  velocity is negative and the control cylinder is thus a source of thrust, not drag.  
 489 On the other hand, the outcome in the upstream region is sensitive to the force model used to  
 490 mimic the effect of the control cylinder, as it turns out its drag balances almost exactly the amount  
 491 by which it reduces the drag of the main cylinder. The weak upstream control efficiency may thus  
 492 be due to the fact that the simplified adjoint method considers only the mean component of the  
 493 force acting on the control cylinder, but overlooks the potential for additional drag reduction via  
 494 the fluctuating component. Moreover, this is a region where the control cylinder likely induces  
 495 strong mean flow modifications because the local inhomogeneity length scale becomes smaller than  
 496 the diameter of the control cylinder, which in turn may invalidate the linear assumption inherent  
 497 to the adjoint method (the retained diameter  $d = 0.1$  is a compromise between smallness and cost  
 498 control, as implementing a smaller control cylinder would require increasing the number of grid  
 499 points and decreasing the time-step to capture properly the wake of the control cylinder).

#### 500 4. Turbulent regime and square geometry at $Re=22000$

501 In order to push the comparison further, additional calculations have been undertaken for a  
 502 square geometry of the main cylinder, whose larger upstream sensitivity yields more clear-cut  
 503 control efficiency, as can be inferred from the results in [25, 92]. This is because the blunt square  
 504 geometry strengthens the upstream pressure gradient (compared to its bluff circular shape). In  
 505 return, the gap flow velocity between the two cylinders decreases and so does the drag of the control  
 506 cylinder, hence a boost in efficiency that helps mitigate the issue of sensitivity to the force model.

507 Another 40 episodes have been run for this case, which represents 320 simulations, each of which  
 508 lasts  $\sim 3\text{h}20$  on 12 cores, hence  $\sim 1020\text{h}$  of total CPU cost (equivalently,  $\sim 130\text{h}$  of resolution  
 509 time). One difficulty for this case is that the main and control cylinders can intersect each other  
 510 under mapping (7), in which case it has been found relevant to simply discard the CFD and  
 511 force the reward to its uncontrolled value. The moving average reward plateaus after about 30  
 512 episodes in figure 12(a), with the optimal drag  $\bar{c}_x^* = 1.49 \pm 0.01$  computed as the average over  
 513 the 5 latest episodes representing a reduction by 30% with respect to the uncontrolled value 2.16  
 514 (close to the reference 2.1 – 2.2 from the literature [96, 97]). The center position of the control  
 515 cylinder exhibits a similarly converging behavior in figures 12(b-c) with  $x_c^* = -2.04 \pm 0.02$ , and  
 516  $y_c^* = 0$ , which suggests that the drag functional has a well-defined global minimum. This is in  
 517 excellent agreement with [95] reporting experimental reduction of the total drag by 30% inserting  
 518 control cylinders of comparable sizes upstream of the main cylinder at a slightly different Reynolds  
 519 number  $Re = 32000$  (the optimal reported position for  $d = 0.1$  being  $x_c^* \sim -2.0$ ). This is also  
 520 in line with the theoretical drag variations computed from the same simplified adjoint method as  
 521 in section IV A 3, whose negative iso-values mapped in figure 12(d) are reproduced from [25]. The  
 522 latter unveil a main region of interest, that extends upstream over approximately 4 diameters, and

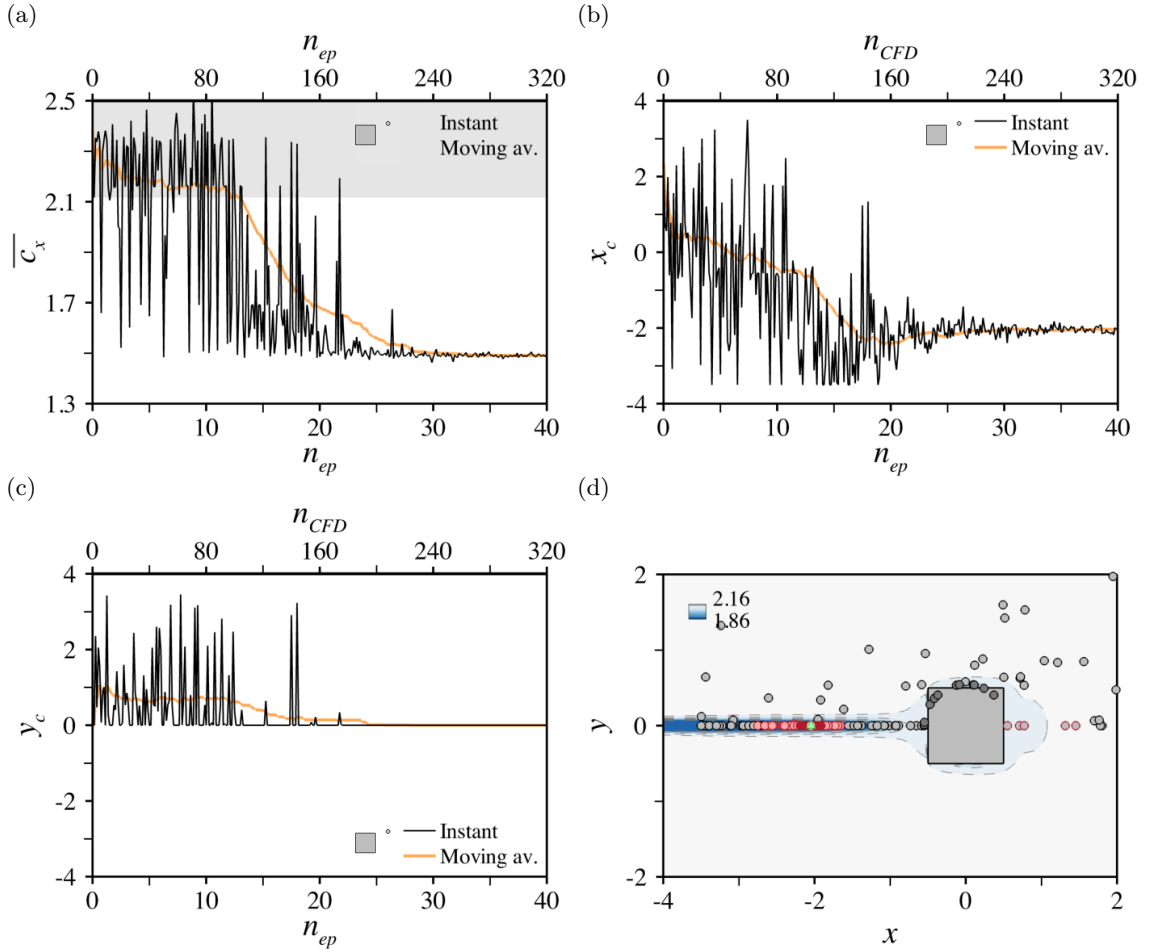


Figure 12. Same as figure 11 for the open-loop control of the square cylinder flow by a small control cylinder of diameter  $d = 0.1$  at  $Re = 22000$ . In (d), the grey symbols circled in red are downstream sub-optimal position. The dark circles are dismissed positions for which the control and main cylinders intersect, and the green triangle marks the experimental position found in [95] to optimally reduce the drag of the two-cylinder system at  $Re = 32000$ .

523 in which drag is reduced by almost 20%, which represents a satisfactory qualitative and quantitative  
 524 compliance with the present PPO-1 results. Drag is also reduced in a second region originating  
 525 from the separation points (pinned here at the front edges), that extends downstream and along  
 526 the outer boundary of the mean recirculation over 1 diameter (similar to what has been found using  
 527 a circular geometry of the main cylinder). It is worth noticing that the algorithm does identify  
 528 sub-optimal positions in this region (shown in figure 12(d) as the grey symbols circled in red). Also,  
 529 a couple of other low-efficiency PPO-1 positions lie further downstream, which is consistent with  
 530 the idea that the simplified adjoint method may miss on additional drag reduction occurring via  
 531 the formation of Reynolds stresses (this is not true of the upstream drag reduction region, whose  
 532 flow is essentially steady, except for low-amplitude oscillations in the gap flow between the two  
 533 cylinders).

## 534 B. Optimal drag reduction of a triangular bluff-body using rotating cylinders

535 The second control problem presented in figure 13(a) is the fluidic pinball [98], an equilateral  
 536 triangle arrangement of three identical circular cylinders oriented against a uniform stream (i.e.,  
 537 the leftmost triangle vertex points upstream, and the rightmost side is orthogonal to the on-coming  
 538 flow), controlled open-loop via user-defined angular velocities. The origin of the coordinate system  
 539 is between the top and bottom cylinders, where we refer to the upstream and downstream cylinders

●	$\overline{c_x}$ $x_c^a$ $y_c$			$\overline{c_x}$ $x_c^a$ $y_c$			$\overline{c_x}$ $x_c^a$ $y_c$			■	Optimal	
	1.51	-1.29	0	1.30	1.72	0	1.46	0.68	0			1.49
	1.54	1.50	0				1.52	-2.40	0			
CFD												
		40			100			3900			22000	Reynolds number
		0.1			»			0.05			»	Time step
		[150; 150]			[100; 200]			»			»	Averaging time span
		$[-15; 40] \times [-15; 15]$			»			»		$[-6; 15] \times [-7; 7]$		Mesh dimensions
		150000			»			»		190000		Nb. mesh elements
		0.001			»			»		»		Interface $\perp$ mesh size
		12			»			»		»		Nb. Cores
PPO-1												
		100			40			»			»	Nb. episodes
		8			»			»			»	Nb. environments
		32			»			»			»	Nb. epochs
		1			»			2			»	Size of mini-batches
		480h			320h			800h			1020h	CPU time
		60h			40h			100h			130h	Resolution time

<sup>a</sup> Only the median value of the optimal interval is reported to ease the presentation.

Table III. Open-loop control of circular and square cylinder flows by a small control cylinder of diameter  $d = 0.1$  - Simulation parameters and convergence data. The interface mesh size yields  $\sim 25 - 40$  grid points in the boundary-layer of the control cylinder, just prior to separation, and the averaging time-span in unsteady flow regimes represents  $\sim 15 - 25$  shedding cycles, depending on the geometry of the main cylinder, the position of the control cylinder and the Reynolds number.

540 as “front”, “top”, and “bottom”, respectively (also labeled 1, 2 and 3 to ease the notation). The  
541 gap spacing  $G = 1.5$  between cylinders yields a master cross-section of 2.5. A turbulent case at  
542  $\text{Re} = U_\infty D / \nu = 2200$  is modeled after the negative Spalart–Allmaras uRANS equations, where  
543  $D$  is the diameter of either cylinder. The objective is to minimize the mean drag  $\overline{D}$  of the three-  
544 cylinder system, using the cylinders individual angular velocities  $\Omega_{1-3}$  as control parameters (with  
545 the convention that  $\Omega_k < 0$  for clockwise rotation). This is a versatile experiment well suited to  
546 challenge the single-step approach, as the requirement to span large ranges of control parameters  
547 emulating a variety of steady and unsteady actuation (e.g., base bleed, suction) under turbulent  
548 conditions makes it especially challenging to rely on the adjoint method (as further discussed in  
549 section V), not to mention DNS.

550

### 1. Steady actuation

551 First, constant angular velocities are applied to each cylinder to alter the vorticity flux fed to the  
552 wake, as evidenced in figure 13(b-d) showing instantaneous vorticity fields computed under several  
553 control configurations. Drag is optimized by minimizing the compound reward function

$$r = -\overline{D} - \beta \sum_{k=1}^3 |\Omega_k|^3, \quad (8)$$

554 where the leftmost term is the power of the drag force and is thus associated to performance, the  
555 rightmost term estimates the power to be supplied to the rotating cylinders and is thus associated to  
556 cost, and  $\beta$  is a weighting coefficient set empirically to  $\beta = 0.025$  (a value found to be large enough  
557 for cost considerations to impact the optimization procedure, but not so large as to dominate  
558 the reward signal, in which case actuating is meaningless). For each PPO-1 learning episode, the  
559 network outputs three values  $\xi_{1-3}$  in  $[-1; 1]^3$  mapped into

$$\Omega_k = \xi_k \Omega_{\max}, \quad (9)$$

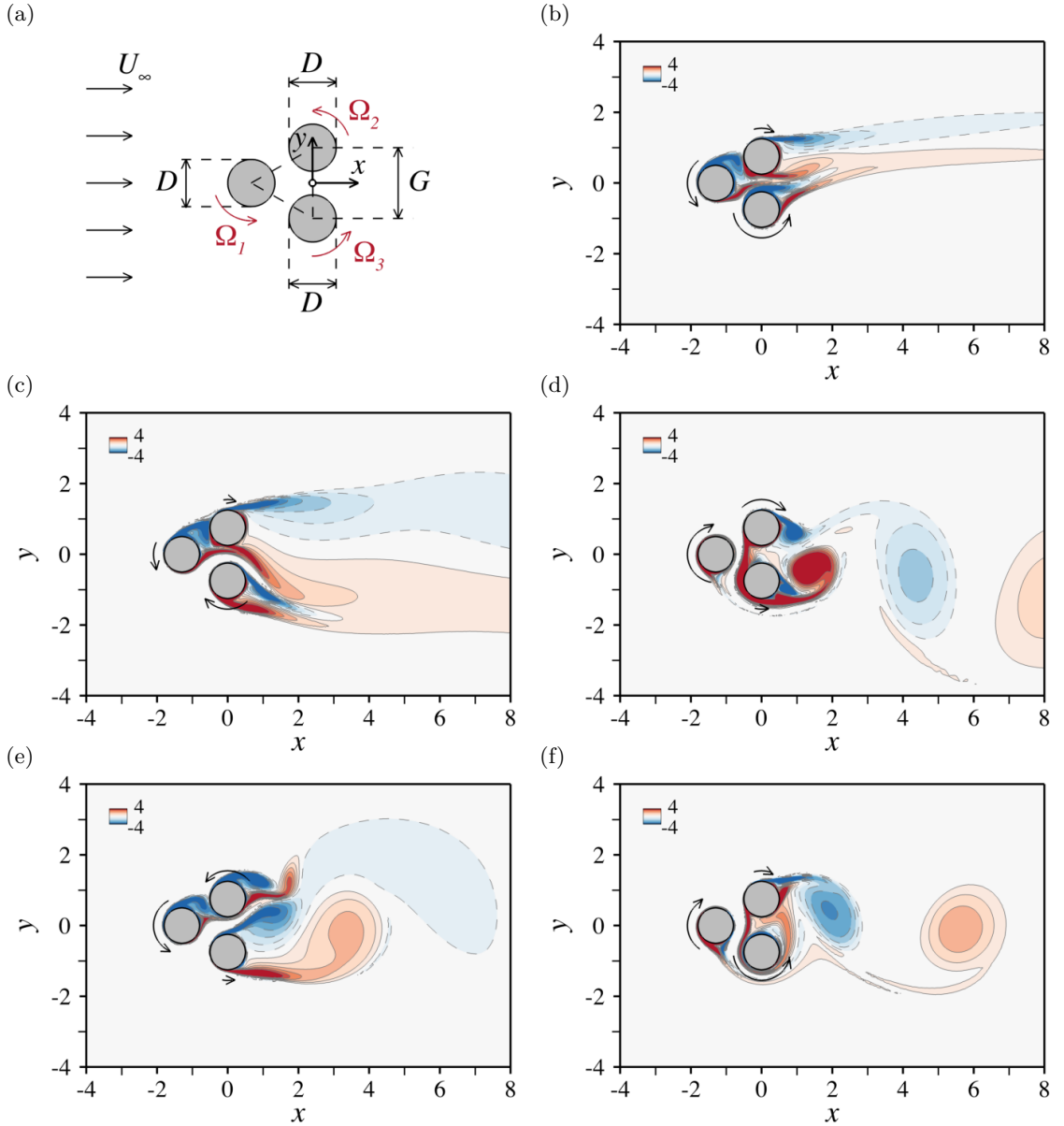


Figure 13. Open-loop control of a fluid pinball - (a) Schematic diagram of the configuration. (b-f) Isocontours of the vorticity field computed at  $\text{Re} = 2200$  for steady angular velocities  $(\Omega_1, \Omega_2, \Omega_3)$  of the individual cylinders, namely (b)  $(3.09, -1.05, 5.00)$ , (c)  $(1.46, -0.62, -2.75)$  (d)  $(-5.00, -2.74, 0.81)$ , (e)  $(3.70, 3.00, 0.61)$  and (f)  $(-3.48, -1.04, 5.00)$ . The rotation directions are marked by the various arrows whose length is proportional to the angular velocity.

560 for the non-dimensional angular velocities to vary in  $[-\Omega_{\max}; \Omega_{\max}]$  with  $\Omega_{\max} = 5$ . The reward  
 561 defined in (8) is computed using the simulation parameters documented in table IV, after which the  
 562 network is updated for 32 epochs using 8 environments and 2 steps mini-batches. Note, rotation  
 563 is actually ramped up over a time-span  $[t_{\Omega_i}; t_{\Omega_f}]$  to smooth out the transient, using effective rates

$$\tilde{\Omega}_k(t) = \frac{\min(\max(t, t_{\Omega_i}), t_{\Omega_f}) - t_{\Omega_i}}{t_{\Omega_f} - t_{\Omega_i}} \Omega_k, \quad (10)$$

564 forced to zero on  $[0, t_{\Omega_i}]$ , to  $\Omega_k$  on  $[t_{\Omega_f}; t_f]$ , and linearly increasing in between.

565 For this case, 120 episodes have been run, which represents 960 simulations, each of which lasts  
 566  $\sim 3\text{h}20$  on 12 cores, hence  $\sim 3200\text{h}$  of total CPU cost (equivalently,  $\sim 400\text{h}$  of resolution time).  
 567 The moving average reward reaches a plateau after about 80 episodes in figure 14(a), where the

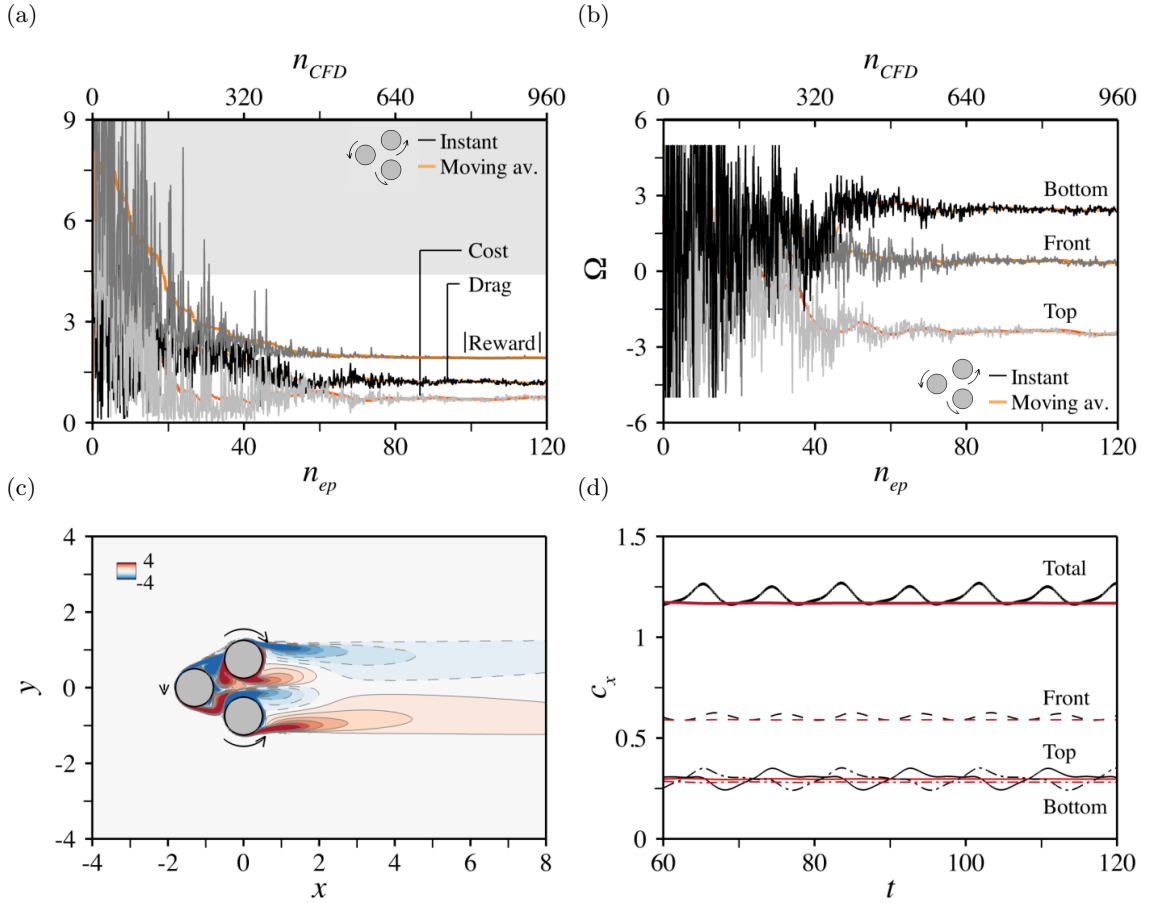


Figure 14. Open-loop control of a fluid pinball at  $Re = 2200$  - (a) Evolution per episode for the instant (black line) and moving average (over episodes, light orange line) values of the mean drag (over time), together with related cost (light grey/dark orange) and reward (dark grey/orange) information computed under steady actuation for  $\beta = 0.025$ . The uncontrolled drag is at the bottom of the grey shaded area. (b) Same as (a) for the angular velocities of the front (black/light orange), top (dark grey/orange) and bottom cylinders (light grey/dark orange). (c) Iso-contours of the vorticity field computed under the optimal velocities  $(\Omega_1^*, \Omega_2^*, \Omega_3^*) = (0.34, -2.49, 2.44)$ . (d) Time history of drag computed under the sub-optimal velocities  $(\Omega_1, \Omega_2, \Omega_3) = (0, -2.47, 2.47)$  (black lines), whose cost is identical to that of the optimal (red lines). The thick lines denote the drag of the three-cylinder system. The thin lines pertain to the front (dashed lines), top (solid lines) and bottom cylinders (dash-dotted lines).

568 relevance of the weighing coefficient value  $\beta = 0.025$  shows through the fact that the performance  
 569 and cost components of the reward are of the same order of magnitude. The optimal value of drag  
 570  $\bar{C}_x^* = 1.17 \pm 0.01$  computed as the average over the 5 latest episodes represents a tremendous  
 571 reduction by almost 60% with respect to the uncontrolled value 2.91. The associated angular  
 572 velocities whose evolution is depicted in figure 14(b) correspond to a boat tail-like arrangement,  
 573 i.e., the top cylinder rotates clockwise ( $\Omega_2^* = -2.49 \pm 0.01$ ), the bottom cylinder rotates counter-  
 574 clockwise and almost symmetrically ( $\Omega_3^* = 2.44 \pm 0.01$ ), and the front cylinder rotates more slowly  
 575 and also counter-clockwise ( $\Omega_1^* = 0.34 \pm 0.01$ ). The net rotation is thus in the same direction as  
 576 the front cylinder, and we show in figure 14(c) that the tilting of the shear layers to the centerline  
 577 alleviates the secondary flow from the gap between the two downstream cylinders, which is found  
 578 to eventually suppress vortex shedding. Interestingly, an experimentally implemented machine  
 579 learning approach using genetic algorithms yields similar optimal arrangements in [34]. For two  
 580 different values of the weighing parameter, the authors therein report optimal angular velocities  
 581  $(0.68, -2.26, 2.56)$  and  $(1.40, -1.70, 2.04)$  and optimal drag reductions by 78% and 49%, respectively,  
 582 but it is uneasy to push further the comparison because the latter study uses a different reward  
 583 function in which drag is approximated from a small, discrete number of sensors distributed in the  
 584 wake.

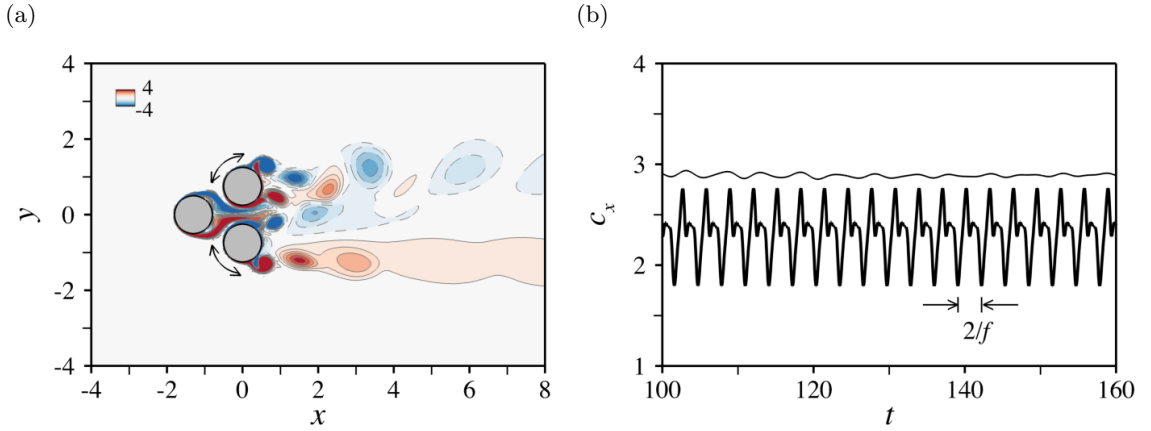


Figure 15. Open-loop control of a fluid pinball at  $Re = 2200$  - (a) Iso-contours of the vorticity field and (b) time history of drag computed under periodic actuation (11) with angular velocity  $\Omega = 2.47$  and frequency  $f = 4f_0$ . The thick and fine lines in (b) denote the controlled and uncontrolled values, respectively.

585 For the purpose of reducing drag, the above asymmetrical boat tailing actuation turns to be  
 586 more efficient than its pure, symmetrical counterpart emulated by  $(\Omega_1, \Omega_2, \Omega_3) = (0, -|\Omega|, |\Omega|)$ .<sup>4</sup>  
 587 This is illustrated in figure 14(d) comparing the optimal drag to its symmetrical value computed  
 588 with  $|\Omega| = 2.47$  (to maintain the same cost efficiency, the associated drag reduction being by  
 589  $\sim 58\%$ ). Pure boat tailing is insufficient to inhibit vortex shedding, as the symmetrical drag of  
 590 all three individual cylinders is seen to exhibit small but finite-amplitude oscillations. Moreover,  
 591 the drag of the downstream cylinders turns to be roughly identical on average. This suggests that  
 592 the edge of asymmetrical over symmetrical boat tailing lies in its ability to reduce the drag of the  
 593 front cylinder, an effect similar to that of suppressing vortex development and reducing drag by  
 594 creating circulation around a single rotating bluff body [99]. Asymmetrical boat tailing is also more  
 595 efficient than base bleed, another method widely used to reduce drag by blowing fluid directly into  
 596 the wake, and that can be emulated by  $(\Omega_1, \Omega_2, \Omega_3) = (0, |\Omega|, -|\Omega|)$  for the reverse rotation of the  
 597 downstream cylinders to conversely enhance the gap flow in between them (not shown here).

598

## 2. Periodic actuation

599 Periodic actuation at frequency  $f$  has also been considered using a simplified configuration

$$\Omega_1 = 0, \quad \Omega_2 = -\Omega_3 = \Omega \sin(2\pi ft), \quad (11)$$

600 whose front cylinder is fixed, and whose downstream cylinders are periodically and symmetrically  
 601 driven with maximum angular velocity  $\Omega$ . Such a control oscillates between symmetrical boat  
 602 tailing (found to be nearly-optimal under steady actuation) and base-bleed, and we assess the  
 603 extent to which an additional degree of freedom (the oscillation frequency) creates room to improve  
 604 the performance. The optimization relies on the compound reward

$$r = -\bar{D} - 2\beta|\Omega|^3, \quad (12)$$

605 computed using the same weighing parameter  $\beta = 0.025$  as before. For each PPO-1 learning  
 606 episode, the network outputs two values  $\xi_{1,2}$  in  $[-1; 1]^2$  mapped into

$$\Omega = \frac{1 + \xi_1}{2} \Omega_{\max}, \quad \frac{f}{f_0} = \frac{1 - \xi_2}{2} \lambda_{\min} + \frac{1 + \xi_2}{2} \lambda_{\max}, \quad (13)$$

607 where  $f_0 = 0.16$  is the dominant frequency of vortex shedding computed in the absence of control.  
 608 The angular velocity therefore varies in  $[0; \Omega_{\max}]$  with  $\Omega_{\max} = 5$  (the case  $\Omega < 0$  is covered by

<sup>4</sup> At least if  $\beta$  is large enough for cost to matter in the optimization procedure, otherwise the algorithm has been found to converge to the symmetrical boat tailing configuration  $(0, -\Omega_{\max}, \Omega_{\max})$ , and the reverse flow is completely suppressed (not shown here).



	$r$	$\overline{c_x}$	$\Omega_1$	$\Omega_2$	$\Omega_3$		$r$	$\overline{c_x}$	$\Omega$	$f$	
	-1.93	1.17	0.34	-2.47	2.44		-2.91	2.91	0	N/D	Optimal
CFD											
			2200								» Reynolds number
			Steady						Periodic		» Actuation
			0.05						0.025		» Time-step
			[5; 10]								» Rotation ramp-up time span
			[300; 400]								» Averaging time span
			$[-6; 20] \times [-6; 6]$								» Mesh dimensions
			110000								» Nb. mesh elements
			0.001								» Interface $\perp$ mesh size
			12								» Nb. Cores
PPO-1											
			120						40		Nb. episodes
			8								» Nb. environments
			32								» Nb. epochs
			2								» Size of mini-batches
			3200h						2100h		CPU time
			400h						260h		Resolution time

Table IV. Simulation parameters and convergence data for open-loop control of a fluid pinball at  $Re = 2200$ . The interface mesh size yields  $\sim 20$  grid points in the boundary-layer of the non-rotating front, top and bottom cylinder, just prior to separation, and the averaging time-span represents  $\sim 15 - 20$  shedding cycles, depending on the angular velocities. For the periodic case, the time step yields  $\sim 60$  data points over the smallest actuation period.

609 periodicity) and the frequency ratio varies in  $[\lambda_{\min}; \lambda_{\max}]$  with  $\lambda_{\min} = 0.5$  and  $\lambda_{\max} = 4$ . This  
610 is a compromise between size of the parameter space and cost control, as investigating smaller  
611 frequencies would require to increase the averaging time-span, and resolving accurately larger  
612 frequencies would require to decrease the time-step. We shall not go into the details of the obtained  
613 results, because the frequency ratio ends up oscillating randomly in  $[\lambda_{\min}; \lambda_{\max}]$ , while the angular  
614 velocity converges to  $\Omega^* = 0$ . It is definitively possible to reduce drag under the considered periodic  
615 actuation, as we show for instance in figure 15 that a velocity  $\Omega = 2.47$  (identical to that used  
616 previously to compare asymmetrical and symmetrical boat tailing) and a frequency ratio  $\lambda = 4$   
617 reduce drag by 20%, but the cost of doing so is too large, as the associated reward actually increases  
618 by 5% (note the period doubling bifurcation phenomenon in figure 15(b): drag is found to exhibit  
619 sub-harmonic oscillations at half the forcing frequency, which is a classical dynamical responses of  
620 harmonically forced nonlinear oscillators). These are only preliminary results intended to compare  
621 the efficiency of steady and periodic strategies using identical reward functions. We therefore defer  
622 to future work the computation of non-trivial periodic optimal distributions, for which it may be  
623 necessary to modify the reward function and/or to reduce the cost (by adequately decreasing the  
624 weighing parameter).

## 625 V. DISCUSSION

626 This section is intended to provide insight into the efficiency of the single-step PPO algorithm  
627 compared to that of other well-established methods. We skip voluntarily DNS, as systematical op-  
628 timization procedures are useless if a problem is simple enough that a small number of numerical  
629 simulations suffices to find the optimal. This is true of the optimization cases documented in sec-  
630 tion III, although the results remain valuable to assess accuracy and highlight the limit of applying  
631 conservative policy updates to optimize sharp reward functions (that are common occurrence in  
632 low to moderate-Reynolds-number-fluid mechanical systems sustaining linear instabilities).



## A. Adjoint methods

We begin with the adjoint method used in section IV A for systematic validation purposes. As explained in appendix A, this is an approach intended to compute the drag of a control-induced disturbance modeled after the linearized governing equations forced by small-amplitude momentum source  $\delta \mathbf{f}$  and wall velocity  $\delta \mathbf{u}_w$ , without ever computing the disturbance itself. The main assumptions and limitations at various levels of sophistication are reviewed in the appendix, so the line of thought is to describe only the specifics of the control problems considered herein. The general picture is that the baseline adjoint method is accurate and fairly efficient in terms of CPU cost, but demanding in terms of storage and increasingly difficult to apply rigorously when turbulence sets in (this is discussed in appendix A 1). On the other hand, the frozen Reynolds stresses approximation has marginal CPU and storage costs, it carries over to any turbulence modeling under the so-called frozen viscosity assumption, but accuracy must be assessed on a case-by-case basis (see appendix A 2).

### 1. Open-loop control by a small control cylinder

Open-loop control by a small control cylinder is a favorable case in the sense that only the center position of the control cylinder (not its shape, nor its size) is optimized, hence the adjoint problem needs be solved only once. Nonetheless, it comes with a substantial modeling component, as the source term  $\delta \mathbf{f}$  used in the adjoint calculations must adequately represent the effect of a true control cylinder. We use here the pointwise reacting force proposed in [25], equal and opposite to the force felt by a control cylinder of same diameter in a uniform flow at the local, mean velocity. The latter is carefully crafted to reference data, but there are inherent approximations associated with overlooking the lift component of the force induced by the local velocity gradient (since the control cylinder, albeit small, has finite size) and inertia (for the model force at each time instant to be the force that would act if the upstream flow at the same instant was a steady one). This can hurt accuracy and undermine the results in flow regions where the control cylinder drag is close to balancing the decrease in the drag of the main cylinder, all the more so in turbulent regimes where additional simplifications are needed to allow implementing the adjoint method itself (e.g., frozen eddy viscosity and/or Reynolds stresses).

In terms of pure performance, the baseline adjoint method is beyond compare for the laminar, steady case at  $\text{Re} = 40$ , because it merely requires solving a couple of steady solutions (one nonlinear, one linear), and PPO-1 would need converge in less than two episodes to approach that cost. Regarding the laminar, time-dependent case at  $\text{Re} = 100$ , the results reported herein rely on a naive implementation of the adjoint method: all time steps of the uncontrolled solution are written to disk, the adjoint equations are solved over the same time interval and with the same time step, and meaningful time averages of the adjoint-based integrands are computing after discarding the early and late time steps (corresponding to transients of the uncontrolled and adjoint solutions). In practice, this takes 45 Gb of storage. The cost of tackling similarly a three-dimensional (3-D) case with 40 points distributed in the span-wise direction would thus be about 2 Tb (as estimated by simple cross-multiplication), which is close to intractable without sophisticated integration, interpolation and/or checkpointing schemes. Meanwhile, the storage cost of PPO-1 is barely a few hundred Mb overall, and is expected to jump to a few ten Gb in 3-D without any additional development. As for CPU cost, the adjoint method amounts to roughly 7-8 episodes, which is about thrice as less as the number of episodes needed to achieve convergence with PPO-1 (this is an estimation for two numerical simulations oversized by the repeated IO calls, although an exact comparison is difficult because our DRL and adjoint results have been obtained using a different finite element codes on different hardware resources). Finally, for the turbulent cases at  $\text{Re} = 3900$  and  $\text{Re} = 22000$ , the cost of the adjoint method is again marginal, as we relied on the frozen Reynolds stresses formulation for which it suffices to compute a nonlinear uncontrolled mean flow and a linear steady adjoint solution. PPO-1 would need to converge in one single episode to match the cost, but we believe the case at  $\text{Re} = 3900$  to provide clear evidence that the simplifying assumptions can make it intricate to compare both qualitatively and quantitatively.

## 2. Open-loop control of a fluidic pinball

The adjoint modeling of the fluidic pinball is straightforward, since the wall velocity  $\delta \mathbf{u}_w$  is simply the cylinder linear velocity. The challenge for this case rather lies in the large value of the optimal angular velocities (found to induce velocities close to the ambient velocity in the vicinity of the downstream cylinders), that suffice to invalidate the linearity assumption inherent to the adjoint method. On paper, this problem can still be tackled with a nonlinear steepest descent algorithm recursively solving an adjoint problem and modifying the control parameters in the direction of the negative gradient. While it usually takes about ten iterations for fluid mechanical systems to converge (provided relevant update strategy and descent step are used), we did not attempt to do so, as it would magnify the limitations of the adjoint method underlined in the appendix. Namely, the storage cost would increase (even a simple conjugate gradient algorithm would require availability of multiple time histories of adjoint solutions) and convergence could be weakened or even sapped if the simplifications made in turbulent regimes yield inaccurate gradient evaluations.

### B. Evolution strategies

Evolution strategies (ES) are another popular family of division of population-based algorithms performing black-box optimization in continuous search spaces without computing directly the gradient of the target function. ES imitate principles of organic evolution processes as rules for optimum seeking procedures, using repeated interplay of variation (via recombination and mutation) and selection in populations of candidate solutions. They rely on a stochastic description of the variables to optimize, i.e., they consider probability density functions instead of deterministic variables. At each generation (or iteration) new candidate solutions are sampled isotropically by variation of the current parental individuals according to a multivariate normal distribution. After applying recombination and mutation transformations (respectively amounting to selecting a new mean for the distribution, and to adding a random perturbation with zero mean), the individuals with the highest cost function are then selected to become the parents in the next generation. Improved variants include the covariance matrix adaptation evolution strategy (CMA-ES), that also updates its full covariance matrix to accelerate convergence toward the optimum (which amounts to learning a second-order model of the underlying objective function).

As has been said for introductory purposes, it lies out of the scope of this paper to provide exhaustive performance comparison data against state-of-the art evolution algorithms. The efforts for developing single-step PPO remain at an early stage, so we do not expect the method to be able to compete right away. Nonetheless, we do not expect it to be utterly outmatched either, as genetic algorithms<sup>5</sup> have been shown capable to learn optimal open- and closed-loop control strategies within a few hundreds to a few thousands test runs (see [100] and the references therein), and it takes a few hundred (resp. less than one thousand) simulations for single-step PPO to learn the optimal open-loop strategy for control by a small cylinder (resp. for control of the fluidic pinball). In present form, the method can be thought as an evolutionary-like algorithm with simpler heuristics (i.e., without an evolutionary update strategy, since the optimal model parameters are learnt via gradient ascent). Its performance should thus be comparable to that of standard ES methods with isotropic covariance matrix, meaning that further characterization and fine-tuning, as well as pre-trained deep learning models (as is done in transfer learning) are likely required to outperform more advanced methods.

## VI. CONCLUSION

Open-loop control of laminar and turbulent flow past bluff bodies is achieved here training a fully connected network with a novel single-step PPO deep reinforcement algorithm, in which it gets only one attempt per learning episode at finding the optimal. The numerical reward fed to

<sup>5</sup> Another class of evolutionary algorithms with slightly different implementation details. Namely, most parameters in genetic algorithms (GA) are exogenous, i.e., set by the practitioner, while ES features endogenous parameters associated with individuals, that evolve together with them. Also, only the fittest individuals are selected to become parents in GA, while parents are selected randomly in ES and the fittest offsprings are selected and inserted in the next generation.

730 the network is computed with a finite elements CFD environment solving stabilized weak forms  
 731 of the governing equations (Navier–Stokes, otherwise uRANS with negative Spalart–Allmaras as  
 732 turbulence model) with a combination of variational multiscale approach, immersed volume method  
 733 and anisotropic mesh adaptation.

734 Convergence and accuracy are assessed from two optimization cases (maximizing the mean lift of  
 735 a NACA 0012 airfoil and the fluctuating lift of two side-by-side circular cylinders, both in laminar  
 736 regimes). Those are simple enough to allow comparison to in-house DNS data, yet they stress  
 737 that the occurrence of instability yields sharp reward functions for which the conservative policy  
 738 updates specific to PPO can trap the optimization process into local optima. The method is also  
 739 applied to two open-loop control problems whose parameter spaces are large enough to dismiss  
 740 DNS. Single-step PPO is found to successfully reduce the drag of laminar and turbulent cylinder  
 741 flows by mapping the best positions for placement of a small control cylinder in good agreement  
 742 with reference data obtained by the adjoint method. The achieved reduction ranges from 2%  
 743 using a circular geometry of the main cylinder at  $Re = 40$ , up to 30% using a square geometry  
 744 at  $Re = 22000$ . Second, the method proves fruitful to reduce the drag of the fluidic pinball, an  
 745 arrangement of three identical, rotating circular cylinders immersed in a turbulent stream. An  
 746 optimal reduction by almost 60% (consistent with that recently obtained using genetic algorithms)  
 747 is reported using a boat tailing actuation made up of a slowly rotating front cylinder and two  
 748 downstream cylinders rotating in opposite directions so as to reduce the gap flow in between them.  
 749 For both cases, convergence is reached after a few ten episodes, which represents a few hundreds  
 750 CFD runs. Exhaustive computational efficiency data are reported with the hope to foster future  
 751 comparisons, but it is worth emphasizing that we did not seek to optimize said efficiency, neither  
 752 by optimizing the hyper parameters, nor by using pre-trained deep learning models.

753 Fluid dynamicists have just begun to gauge the relevance of deep reinforcement learning tech-  
 754 niques to assist the design of optimal flow control strategies. This research weighs in on this issue  
 755 and shows that the proposed single-step PPO holds a high potential as a reliable, go-to black-box  
 756 optimizer for complex CFD problems. The one advantages here are scope and applicability, as the  
 757 storage cost of an episode is simply that of a CFD run (times the number of environments), and  
 758 there is no prerequisite beyond the ability to compute accurate numerical solutions (which behoves  
 759 the CFD solver, not the RL algorithm). Consequently, we would not anticipate any additional  
 760 numerical developments before tackling a 3-D turbulent flow with the same CFD environment,  
 761 even with a more sophisticated turbulence modeling (since the built-in small-scale component of  
 762 the VMS solution also acts as an implicit LES). Despite these achievements, further development,  
 763 characterization and fine-tuning are needed to consolidate the acquired knowledge, whether it be  
 764 via an improved balance between exploration and exploitation to deal with steep global maxima  
 765 (for instance using Trust Region-Guided PPO, as it effectively encourages the policy to explore  
 766 more on the potential valuable actions, no matter whether they were preferred by the previous  
 767 policies or not), via non-normal probability density functions to deal with multiple global maxima,  
 768 or via coupling with a surrogate model trained on-the-fly.

## 769 Appendix A: A quick survey of adjoint-based optimization

770 We briefly review here the various adjoint frameworks used in section IV A for systematic vali-  
 771 dation purposes of the PPO-1 results. The starting point is a so-called uncontrolled solution  $(\mathbf{u}, p)$   
 772 to the non-linear equations of motion (Navier–Stokes, unless specified otherwise) forced by a mo-  
 773 mentum source  $\mathbf{f}$  and a velocity  $\mathbf{u}_w$  distributed over all solid surfaces  $\Gamma_w$  in the computational  
 774 domain (although it is possible to restrict to a subset).

### 775 1. Baseline adjoint method

776 The adjoint method computes the change in drag induced by small variations  $(\delta\mathbf{f}, \delta\mathbf{u}_w)$  of these  
 777 control parameters as

$$\delta\bar{c}_x = \int_{\Omega} \overline{\mathbf{u}^\dagger \cdot \delta\mathbf{f}} \, ds + \int_{\Gamma_w} \overline{(\boldsymbol{\sigma}^\dagger(-p^\dagger, \mathbf{u}^\dagger) \cdot \mathbf{n}) \cdot \delta\mathbf{u}_w} \, dl, \quad (\text{A1})$$

778 where  $\mathbf{n}$  is the unit outward normal to  $\Gamma_w$  and we note  $\boldsymbol{\sigma}^\dagger(-p^\dagger, \mathbf{u}^\dagger) = p^\dagger \mathbf{I} + \frac{1}{\text{Re}} \nabla \mathbf{u}^\dagger$ . Finally,  
779  $(\mathbf{u}^\dagger, p^\dagger)$  are adjoint velocity and pressure fields solution to

$$\nabla \cdot \mathbf{u}^\dagger = 0, \quad -(\partial_t \mathbf{u}^\dagger + \nabla \mathbf{u}^\dagger \cdot \mathbf{u}) + \nabla \mathbf{u}^T \cdot \mathbf{u}^\dagger + \nabla \cdot \boldsymbol{\sigma}^\dagger(-p^\dagger, \mathbf{u}^\dagger) = \mathbf{0}, \quad (\text{A2})$$

780 forced at  $\Gamma_w$  by a velocity equal to twice the ambient velocity (the factor of 2 stems from the  
781 definition of dynamic pressure), as obtained multiplying  $\mathbf{u}^\dagger$  and  $p^\dagger$  onto the linearized momentum  
782 and continuity equations, using the divergence theorem to integrate by parts over the computational  
783 domain, and integrating in time over the span of the simulation. In essence, this amounts to  
784 computing the drag of the control induced disturbance modeled after the forced, linearized Navier–  
785 Stokes equations, without ever computing the disturbance itself.

786 A typical implementation consists of two sequential numerical simulations (for the uncontrolled  
787 and adjoint solutions, respectively) plus a series of vector dot products, to give the drag variation  
788 at each grid point. This is simple on paper, but the method has some limitations :

789 - the adjoint equations are problem-specific and must be derived and implemented manually  
790 on a case-by-case basis.

791 - the cost is marginal in steady flow regimes, because the time-independence of the uncontrolled  
792 solution makes the adjoint problem purely linear. Otherwise, the entire time history of uncontrolled  
793 solutions must be available at every adjoint time step because of the reversal of space-time direc-  
794 tionality; see the minus sign ahead of the material derivative term in eqs. (A2). This is very  
795 demanding in terms of storage (the repeated IO also increases the computational burden compared  
796 to a classical CFD run with identical simulation parameters) but these issues can be mitigated  
797 using checkpointing [101] and high-order time-integration and interpolation schemes [102].

798 - not all cost functions are admissible due to the need for consistent adjoint boundary conditions,  
799 although this can be overcome with augmented Lagrangian methods based on auxiliary boundary  
800 equations [103].

801 - applicability to high-fidelity turbulence modeling is uncertain because the noise-induced sen-  
802 sitivity to initial conditions (the “butterfly effect”) is expected to yield exponentially diverging  
803 solutions if the length of the adjoint simulation exceeds the predictability time scale. Possible  
804 solutions include averaging over a large number of ensemble calculations [104] (which increases  
805 significantly the computational cost and decreases the attractiveness of the method) or invoking  
806 sophisticated shadowing and space-split techniques sampling on selected flow trajectories [105, 106]  
807 (which comes at the cost of ease of implementation). Moreover, the literature somehow oddly re-  
808 ports several cases of turbulent adjoint solution blowing up in 2-D [107, 108] and 3-D [109], but  
809 also several instances in 3-D where no blow-up is observed [110–112].

810 - applicability to RANS simulations is conversely generally acknowledged. However, discarding  
811 the linearization and adjointization of even the simplest turbulence models (using the so-called  
812 frozen eddy-viscosity approximation) to avoid massive debugging and validation efforts has some-  
813 how become standard lore, even though completeness and exactness are required to ensure numer-  
814 ical accuracy and avoid diverging adjoint solutions due to error propagation and amplification.

## 815 2. Frozen Reynolds stresses approximation

816 A simple adjoint formalism has been proposed in [25] to provide insight into the reliability of  
817 adjoint-based predictions in practical situations where no complete history of time and space-  
818 accurate solutions is available. The approach is closely related to existing studies considering the  
819 mean flow an admissible solution for linear stability analysis, as it simply dismisses the way the  
820 control-induced modification to the fluctuating uncontrolled solution feeds back onto the mean  
821 (hence the frozen Reynolds stress moniker to echo the above frozen eddy viscosity). In doing so,  
822 (A1) can be shown to reduce to

$$\delta \overline{\overline{c_x}} = \int_{\Omega} \overline{\overline{\mathbf{u}^\dagger}} \cdot \delta \overline{\overline{\mathbf{f}}} \, ds + \int_{\Gamma_s} (\boldsymbol{\sigma}^\dagger(\overline{\overline{p^\dagger}}, \overline{\overline{\mathbf{u}^\dagger}}) \cdot \mathbf{n}) \cdot \delta \overline{\overline{\mathbf{u}_w}} \, dl, \quad (\text{A3})$$

823 where the double overline denotes approximations to the true time-averaged quantities, and the  
824 adjoint velocity and pressure fields are solution to

$$\nabla \cdot \overline{\overline{\mathbf{u}^\dagger}} = 0, \quad -\nabla \overline{\overline{\mathbf{u}^\dagger}} \cdot \overline{\overline{\mathbf{u}}} + \nabla \overline{\overline{\mathbf{u}^T}} \cdot \overline{\overline{\mathbf{u}^\dagger}} + \nabla \cdot \boldsymbol{\sigma}(-\overline{\overline{p^\dagger}}, \overline{\overline{\mathbf{u}^\dagger}}) = \mathbf{0}, \quad (\text{A4})$$

825 forced at  $\Gamma_w$  by the same velocity equal to twice the ambient velocity. The strength of the approach  
 826 lies in the fact that once the mean uncontrolled solution is known, computing the approximated  
 827 adjoint solution merely requires solving a single linear problem. Accuracy must be assessed on  
 828 a case-by-case basis, but the computational and storage costs of doing so are marginal, and the  
 829 approach carries over to any turbulence modeling method under the frozen viscosity assumption.

### 830 ACKNOWLEDGEMENTS

831 This work is supported by the Carnot M.I.N.E.S. Institute through the M.I.N.D.S. project.

- 
- 832 [1] J. J. Corbett and H. W. Koehler, Updated emissions from ocean shipping, *J. Geophys. Res.* **108**,  
 833 4650 (2003).  
 834 [2] M. R. Khorrami, M. E. Berkman, and M. Choudhari, Unsteady flow computations of a slat with a  
 835 blunt trailing edge, *AIAA J.* **38**, 2050 (2000).  
 836 [3] C. Rowley, T. Colonius, and A. Basu, On self-sustained oscillations in two-dimensional compressible  
 837 flow over rectangular cavities, *J. Fluid Mech.* **455**, 315 (2002).  
 838 [4] J. Knight, Honey, i shrank the lab, *Nature* **418**, 474 (2002).  
 839 [5] N. Syred and J. M. Beér, Combustion in swirling flows: A review, *Combust. Flame* **23**, 143 (1974).  
 840 [6] M. Gad-el Hak, Modern developments in flow control, *Appl. Mech. Rev.* **49**, 365 (1996).  
 841 [7] J. Lumley and P. Blossey, Control of turbulence, *Annu. Rev. Fluid Mech.* **30**, 311 (1998).  
 842 [8] A. Glezer and M. Amitay, Synthetic jets, *Annu. Rev. Fluid Mech.* **34**, 503 (2002).  
 843 [9] S. S. Collis, R. D. Joslin, A. Seifert, and V. Theofilis, Issues in active flow control: theory, control,  
 844 simulation, and experiment, *Prog. Aerosp. Sci.* **40**, 237 (2004).  
 845 [10] J. Kim and T. R. Bewley, A linear systems approach to flow control, *Annu. Rev. Fluid Mech.* **39**,  
 846 383 (2007).  
 847 [11] H. Choi, W.-P. Jeon, and J. Kim, Control of flow over a bluff body, *Annu. Rev. Fluid Mech.* **40**, 113  
 848 (2008).  
 849 [12] T. C. Corke, C. L. Enloe, and S. P. Wilkinson, Dielectric barrier discharge plasma actuators for flow  
 850 control, *Annu. Rev. Fluid Mech.* **42**, 505 (2010).  
 851 [13] L. N. Cattafesta and M. Sheplak, Actuators for active flow control, *Annu. Rev. Fluid Mech.* **43**, 247  
 852 (2011).  
 853 [14] A. Seifert, Boundary layer separation control: Experimental perspective and modeling outlook, *Annu.*  
 854 *Rev. Fluid Mech.* **50**, null (2018).  
 855 [15] M. C. G. Hall, Application of adjoint sensitivity theory to an atmospheric general circulation model,  
 856 *J. Atmospheric Sci.* **43**, 2644 (1986).  
 857 [16] A. Jameson, Aerodynamic design via control theory, *J. Sci. Comput.* **3**, 233 (1998).  
 858 [17] A. Jameson, L. Martinelli, and N. A. Pierce, Fluid dynamics optimum aerodynamic design using the  
 859 Navier–Stokes equations, *Theor. Comput. Fluid Dyn.* **10**, 213 (1998).  
 860 [18] M. D. Gunzburger, *Perspectives in flow control and optimization*, SIAM, Philadelphia (2002).  
 861 [19] B. Mohammadi and O. Pironneau, Shape optimization in fluid mechanics, *Annu. Rev. Fluid Mech.*  
 862 **36**, 255 (2004).  
 863 [20] D. C. Hill, A theoretical approach for analyzing the restabilization of wakes, NASA Technical Mem-  
 864 orandum NASA-TM-103858 (1992).  
 865 [21] F. Giannetti and P. Luchini, Structural sensitivity of the first instability of the cylinder wake, *J.*  
 866 *Fluid Mech.* **581**, 167 (2007).  
 867 [22] O. Marquet, D. Sipp, and L. Jacquin, Sensitivity analysis and passive control of cylinder flow, *J.*  
 868 *Fluid Mech.* **615**, 221 (2008).  
 869 [23] J. O. Pralits, L. Brandt, and F. Giannetti, Instability and sensitivity of the flow around a rotating  
 870 circular cylinder, *J. Fluid Mech.* **650**, 513 (2010).  
 871 [24] D. Sipp, O. Marquet, P. Meliga, and A. Barbagallo, Dynamics and control of global instabilities in  
 872 open-flows: a linearized approach, *Appl. Mech. Rev.* **63**, 030801 (2010).  
 873 [25] P. Meliga, E. Boujo, G. Pujals, and F. Gallaire, Sensitivity of aerodynamic forces in laminar and  
 874 turbulent flow past a square cylinder, *Phys. Fluids* **26**, 104101 (2014).  
 875 [26] A. Krizhevsky, I. Sutskever, and G. E. Hinton, Imagenet classification with deep convolutional neural  
 876 networks, *Proceedings of the 25th International Conference on Neural Information Processing Systems*  
 877 , 1097 (2012).  
 878 [27] J. Kober, J. A. Bagnell, and J. Peters, Reinforcement learning in robotics: A survey, *The International*  
 879 *Journal of Robotics Research* **32**, 1238 (2013).

- 880 [28] V. Mnih, K. Kavukcuoglu, D. Silver, R. A. A., V. J., M. G. Bellemare, A. Graves, M. Riedmiller,  
881 A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran,  
882 D. Wierstra, S. Legg, and D. Hassabis, Human-level control through deep reinforcement learning,  
883 *Nature* **518**, 7540 (2015).
- 884 [29] B. Lusch, J. N. Kutz, and S. L. Brunton, Deep learning for universal linear embeddings of nonlinear  
885 dynamics, *Nature communications* **9**, 1 (2018).
- 886 [30] M. Raissi, A. Yazdani, and G. E. Karniadakis, Hidden Fluid Mechanics: A Navier-Stokes Informed  
887 Deep Learning Framework for Assimilating Flow Visualization Data, *arXiv* (2018).
- 888 [31] A. D. Beck, D. G. Flad, and C.-D. Munz, Deep Neural Networks for Data-Driven Turbulence Models,  
889 *arXiv* (2018).
- 890 [32] Z. Wang, D. Xiao, F. Fang, R. Govindan, C. C. Pain, and Y. Guo, Model identification of reduced  
891 order fluid dynamics systems using deep learning, *Int. J. Numer. Meth. Fluids* **86**, 255 (2018).
- 892 [33] N. Gautier, J.-L. Aider, T. Duriez, and B. R. Noack, Closed-loop separation control using machine  
893 learning, *J. Fluid Mech.* **770**, 442 (2015).
- 894 [34] C. Raibaudo, P. Zhong, B. R. Noack, and R. J. Martinuzzi, Machine learning strategies applied to  
895 the control of a fluidic pinball, *Phys. Fluids* **32**, 015108 (2020).
- 896 [35] Y. Lee, H. Yang, and Z. Yin, PIV-DCNN: cascaded deep convolutional neural networks for particle  
897 image velocimetry, *Exp Fluids* **58**, 171 (2017).
- 898 [36] J. Rabault, J. Kolaas, and A. Jensen, Performing particle image velocimetry using artificial neural  
899 networks: a proof-of-concept, *Meas. Sci. Technol.* **28**, 125301 (2017).
- 900 [37] S. Cai, J. Liang, Q. Gao, C. Xu, and R. Wei, Particle image velocimetry based on a deep learning  
901 motion estimator, *IEEE Trans. Instrum. Meas.* **69**, 3538 (2020).
- 902 [38] S. L. Brunton, B. R. Noack, and P. Koumoutsakos, Machine learning for fluid mechanics, *Annu. Rev.*  
903 *Fluid Mech.* **52**, 477 (2020).
- 904 [39] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker,  
905 M. Lai, A. Bolton, Y. Chen, T. P. Lillicrap, F. Hui, L. Sifre, G. van den Driessche, T. Graepel, and  
906 D. Hassabis, Mastering the game of go without human knowledge, *Nature* **550**, 354 (2017).
- 907 [40] M. Moravčik, M. Schmid, N. Burch, V. Lisy, D. Morrill, N. Bard, T. Davis, K. Waugh, M. Johanson,  
908 and M. Bowling, DeepStack: expert-level artificial intelligence in heads-up no-limit poker, *Science*  
909 **356**, 508 (2017).
- 910 [41] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, Proximal Policy Optimization  
911 Algorithms, *arXiv e-prints*, *arXiv:1707.06347* (2017), *arXiv:1707.06347 [cs.LG]*.
- 912 [42] J. Hwangbo, J. Lee, A. Dosovitskiy, D. Bellicoso, V. Tsounis, V. Koltun, and M. Hutter, Learning  
913 agile and dynamic motor skills for legged robots, *Science Robotics* **4**, eaau5872 (2019).
- 914 [43] A. V. Bernstein and E. V. Burnaev, Reinforcement learning in computer vision, *Proc. SPIE 10696*,  
915 *10th International Conference on Machine Vision (ICMV 2017)* (2018).
- 916 [44] X. Pan, Y. You, Z. Wang, and C. Lu, Virtual to real reinforcement learning for autonomous driving,  
917 *arXiv preprint arXiv:1704.03952*. (2017).
- 918 [45] Y. Deng, F. Bao, Y. Kong, Z. Ren, and Q. Dai, Deep direct reinforcement learning for financial signal  
919 representation and trading, *IEEE Trans. Neural Netw. Learn. Syst.* **28**, 653 (2017).
- 920 [46] X. Y. Lee, A. Balu, D. Stoecklein, B. Ganapathysubramanian, and S. Sarkar, Flow shape design for  
921 microfluidic devices using deep reinforcement learning, *arXiv preprint arXiv:1811.12444* (2018).
- 922 [47] X. Yan, J. Zhu, M. Kuang, and X. Wang, Aerodynamic shape optimization using a novel optimizer  
923 based on machine learning techniques, *Aerosp. Sci. Technol.* **86**, 826 (2019).
- 924 [48] J. Viquerat, J. Rabault, A. Kuhnle, H. Ghraieb, and E. Hachem, Direct shape optimization through  
925 deep reinforcement learning, *arXiv preprint arXiv:1908.09885* (2019).
- 926 [49] P. Ma, Y. Tian, Z. Pan, B. Ren, and D. Manocha, Fluid directed rigid body control using deep  
927 reinforcement learning, *ACM Transactions on Graphics (TOG)* **37**, 1 (2018).
- 928 [50] L. Biferale, F. Bonaccorso, M. Buzicotti, P. Clark Di Leioni, and K. Gustavsson, Zermelo's problem:  
929 Optimal point-to-point navigation in 2D turbulent flows using reinforcement learning, *Chaos* **29**,  
930 103138 (2019).
- 931 [51] F. Ren, H. Hu, and H. Tang, Active flow control using machine learning: A brief review, *J. Hydro-*  
932 *dynam.* **32**, 247 (2020).
- 933 [52] V. Belus, J. Rabault, J. Viquerat, Z. Che, E. Hachem, and U. Réglade, Exploiting locality and  
934 translational invariance to design effective deep reinforcement learning control of the 1-dimensional  
935 unstable falling liquid film, *AIP Adv.* **9**, 125014 (2019).
- 936 [53] M. A. Bucci, O. Semeraro, A. Allauzen, G. Wisniewski, L. Cordier, and L. Mathelin, Control of  
937 chaotic systems by deep reinforcement learning, *Proc. Roy. Soc. A* **475**, 20190351 (2019).
- 938 [54] G. Novati, L. Mahadevan, and P. Koumoutsakos, Controlled gliding and perching through deep-  
939 reinforcement-learning, *Phys. Rev. Fluids* **4**, 093902 (2019).
- 940 [55] G. Novati, S. Verma, D. Alexeev, D. Rossinelli, W. M. van Rees, and P. Koumoutsakos, Synchroni-  
941 sation through learning for two self-propelled swimmers, *Bioinspir. Biomim.* **12**, 036001 (2017).
- 942 [56] S. Verma, G. Novati, and P. Koumoutsakos, Efficient collective swimming by harnessing vortices  
943 through deep reinforcement learning, *Proc. Natl. Acad. Sci. U.S.A.* **115**, 5849 (2018).

- 944 [57] J. Rabault, M. Kuchta, A. Jensen, U. Réglade, and N. Cerardi, Artificial neural networks trained  
945 through deep reinforcement learning discover control strategies for active flow control, *Journal of*  
946 *Fluid Mechanics* **865**, 281 (2019).
- 947 [58] H. Tang, J. Rabault, A. Kuhnle, Y. Wang, and T. Wang, Robust active flow control over a range  
948 of Reynolds numbers using an artificial neural network trained through deep reinforcement learning,  
949 *Phys. Fluids* **32**, 053605 (2020).
- 950 [59] R. Paris, R. Beneddine, and J. Dandois, Robust flow control and optimal sensor placement using  
951 deep reinforcement learning, arXiv preprint arXiv:2006.11005 (2020).
- 952 [60] H. Xu, W. Zhang, J. Deng, and J. Rabault, Active flow control with rotating cylinders by an artificial  
953 neural network trained by deep reinforcement learning, *J. Hydrodyn.* **32**, 254 (2020).
- 954 [61] F. Ren, J. Rabault, and H. Tang, Applying deep reinforcement learning to active flow control in  
955 turbulent conditions, arXiv preprint arXiv:2006.10683 (2020).
- 956 [62] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, Learning representations by back-propagating  
957 errors, *Nature* **323**, 533 (1986).
- 958 [63] Y. Wang, H. He, X. Tan, and Y. Gan, Trust region-guided proximal policy optimization, arXiv  
959 preprint arXiv:1901.10314 (2019).
- 960 [64] T. Coupez and E. Hachem, Solution of high-reynolds incompressible flow with stabilized finite element  
961 and adaptive anisotropic meshing, *Comput. Methods Appl. Mech. Engrg.* **267**, 65 (2013).
- 962 [65] T. J. R. Hughes, G. R. Feijóo, L. Mazzei, and J.-B. Quincy, The variational multiscale method - a  
963 paradigm for computational mechanics, *Comput. Methods Appl. Mech. Engrg.* **166**, 3 (1998).
- 964 [66] R. Codina, Stabilization of incompressibility and convection through orthogonal sub-scales in finite  
965 element methods, *Comput. Methods Appl. Mech. Engrg.* **190**, 1579 (2000).
- 966 [67] Y. Bazilevs, V. M. Calo, J. A. Cottrell, T. J. R. Hughes, A. Reali, and G. Scovazzi, Variational multi-  
967 scale residual-based turbulence modeling for large eddy simulation of incompressible flows, *Comput.*  
968 *Methods Appl. Mech. Engrg.* **197**, 173 (2007).
- 969 [68] S. R. Allmaras, F. T. Johnson, and P. R. Spalart, Modifications and clarifications for the implementa-  
970 tion of the Spalart–Allmaras turbulence model, *Proc. 7th International Conference on Computational*  
971 *Fluid Dynamics - ICCFD8-1902* (2012).
- 972 [69] R. Codina, Comparison of some finite element methods for solving the diffusion-convection-reaction  
973 equation, *Comput. Methods Appl. Mech. Engrg.* **156**, 185 (1998).
- 974 [70] S. Badia and R. Codina, Analysis of a stabilized finite element approximation of the transient  
975 convection-diffusion equation using an ALE framework, *SIAM Journal on Numerical Analysis* **44**,  
976 2159 (2006).
- 977 [71] J. Bruchon, H. Dignonnet, and T. Coupez, Using a signed distance function for the simulation of metal  
978 forming processes: formulation of the contact condition and mesh adaptation, *Int. J. Numer. Meth.*  
979 *Eng.* **78**, 980 (2004).
- 980 [72] C. Gruau and T. Coupez, 3d tetrahedral, unstructured and anisotropic mesh generation with adap-  
981 tation to natural and multidomain metric, *Comput. Methods Appl. Mech. Engrg.* **194**, 4951 (2005).
- 982 [73] E. Hachem, B. Rivaux, T. Kloczko, H. Dignonnet, and T. Coupez, Stabilized finite element method  
983 for incompressible flows with high Reynolds number, *J. Comput. Phys.* **229**, 8643 (2010).
- 984 [74] T. Coupez, G. Jannoun, N. Nassif, H. C. Nguyen, H. Dignonnet, and E. Hachem, Adaptive time-step  
985 with anisotropic meshing for incompressible flows, *J. Comput. Phys.* **241**, 195 (2013).
- 986 [75] J. Sari, F. Cremonesi, M. Khalloufi, F. Cauneau, P. Meliga, Y. Mesri, and E. Hachem, Anisotropic  
987 adaptive stabilized finite element solver for rans models, *Int. J. Numer. Meth. Fluids* **86**, 717 (2018).
- 988 [76] G. Guiza, A. Larcher, A. Goetz, L. Billon, P. Meliga, and E. Hachem, Anisotropic boundary layer  
989 mesh generation for reliable 3D unsteady RANS simulations, *Finite Elem. Anal. Des.* **170**, 103345  
990 (2020).
- 991 [77] E. Hachem, H. Dignonnet, E. Massoni, and T. Coupez, Immersed volume method for solving natural  
992 convection, conduction and radiation of a hat-shaped disk inside a 3d enclosure, *International Journal*  
993 *of numerical methods for heat & fluid flow* **22**, 718 (2012).
- 994 [78] E. Hachem, S. Feghali, R. Codina, and T. Coupez, Immersed stress method for fluid-structure inter-  
995 action using anisotropic mesh adaptation, *Int. J. Numer. Meth. Eng.* **94**, 805 (2013).
- 996 [79] W. Rodi, Comparison of LES and RANS calculations of the flow around bluff bodies, *J. Wind Eng.*  
997 *Ind. Aerodyn.* **69–71**, 55 (1997).
- 998 [80] V. John, Parallele Lösung der inkompressiblen Navier–Stokes Gleichungen auf adaptiv verfeinerten  
999 Gittern, *Otto-von-Guericke-Universität Magdeburg, Fakultät für Mathematik* (1997).
- 1000 [81] V. John, Reference values for drag and lift of a two-dimensional time-dependent ow around a cylinder,  
1001 *Int. J. Numer. Meth. Fluids* **44**, 777 (2004).
- 1002 [82] A. Hill, A. Raffin, M. Ernestus, A. Gleave, A. Kanervisto, R. Traore, P. Dhariwal, C. Hesse,  
1003 O. Klimov, A. Nichol, M. Plappert, A. Radford, J. Schulman, S. Sidor, and Y. Wu, Stable base-  
1004 lines, <https://github.com/hill-a/stable-baselines> (2018).
- 1005 [83] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba,  
1006 Openai gym (2016), arXiv:1606.01540.

- 1007 [84] S. Mittal, V. Kumar, and A. Raghuvanshi, Unsteady incompressible flows past two cylinders in  
1008 tandem and staggered arrangements, *Int. J. Numer. Meth. Fl.* **25**, 1315 (1997).
- 1009 [85] J. R. Meneghini, F. Saltara, C. L. R. Siqueira, and J. A. Ferrari, Numerical simulation of flow  
1010 interference between two circular cylinders in tandem and side-by-side arrangements, *J. Fluids Struct.*  
1011 **15**, 327 (2001).
- 1012 [86] B. Sharman, F.-S. Lien, L. Davidson, and C. Norberg, Numerical predictions of low reynolds number  
1013 flows over two tandem circular cylinders, *Int. J. Numer. Meth. Fl.* **47**, 423 (2005).
- 1014 [87] K. Lee, K.-S. Yang, and D.-H. Yoon, Flow-induced forces on two circular cylinders in proximity,  
1015 *Comput. Fluids* **38**, 111 (2009).
- 1016 [88] X. Mao, Sensitivity of forces to wall transpiration in flow past an aerofoil, *Proc. R. Soc. A* **471**,  
1017 20150618 (2015).
- 1018 [89] P. Meliga, E. Boujo, M. Meldi, and F. Gallaire, Revisiting the drag reduction problem using adjoint-  
1019 based distributed forcing of laminar and turbulent flows over a circular cylinder, *Eur. J. Mech.*  
1020 *B-Fluid* **72**, 123 (2018, accepted).
- 1021 [90] B. Fornberg, A numerical study of steady viscous flow past a circular cylinder, *J. Fluid Mech.* **98**,  
1022 819 (1980).
- 1023 [91] R. D. Henderson, Details of the drag curve near the onset of vortex shedding, *Phys. Fluids* **7**, 2102  
1024 (1995).
- 1025 [92] P. Meliga, Computing the sensitivity of drag and lift in flow past a circular cylinder: time-stepping  
1026 vs. self-consistent analysis, *Phys. Rev. Fluids* **2**, 073905 (2017).
- 1027 [93] H. Sakamoto and H. Haniu, Optimum suppression of fluid forces acting on a circular cylinder, *J.*  
1028 *Fluids Eng.* **116**, 221 (1994).
- 1029 [94] F. Pereira, G. Vaz, and L. Eça, Flow past a circular cylinder: a comparison between rans and hybrid  
1030 turbulence models for a low Reynolds number, *OMAE 2015-41235* (2015).
- 1031 [95] T. Igarashi, Drag reduction of a square prism by flow control using a small rod, *J. Wind Eng. Ind.*  
1032 *Aerodyn.* **69**, 141 (1997).
- 1033 [96] G. Iaccarino, A. Ooi, P. A. Durbin, and M. Behnia, Reynolds averaged simulation of unsteady  
1034 separated flow, *Int. J. Heat Fluid Flow* **24**, 147 (2003).
- 1035 [97] W. Rodi, J. H. Ferziger, M. Breuer, and M. Pourquie, Status of large-eddy simulation: Results of a  
1036 workshop, *J. Fluids Eng.* **119**, 248 (1997).
- 1037 [98] C. Raibaudo, P. Zhong, R. J. Martinuzzi, and B. R. Noack, Open and closed-loop control of a  
1038 triangular bluff body using rotating cylinders, *IFAC-PapersOnLine* **50**, 12291 (2017).
- 1039 [99] S. Kang, H. Choi, and S. Lee, Laminar flow past a rotating circular cylinder, *Phys. Fluids* **11**, 3312  
1040 (1999).
- 1041 [100] Y. Deng, L. Pastur, M. Morzyński, and B. R. Noack, Route to chaos in the fluidic pinball, *Procs. of*  
1042 *the ASME 2018 5th Joint US-European Fluids Engineering Division Summer Meeting* (2018).
- 1043 [101] A. Griewank and A. Walther, An implementation of checkpointing for the reverse or adjoint mode  
1044 of computational differentiation, *ACM T. Math. Software* **26**, 19 (2000).
- 1045 [102] C. Tsitouras, Runge–Kutta pairs of order 5(4) satisfying only the first column simplifying assumption,  
1046 *Comput. Math. with Appl.* **62**, 770 (2011).
- 1047 [103] E. Arian and M. D. Salas, Admitting the inadmissible: adjoint formulation for incomplete cost  
1048 functionals in aerodynamic optimization, *AIAA J.* **37**, 37 (1999).
- 1049 [104] D. J. Lea, M. Allen, and T. W. N. Haines, Sensitivity analysis of the climate of a chaotic system,  
1050 *Tellus A* **52**, 523 (2000).
- 1051 [105] Q. Wang, Forward and adjoint sensitivity computation of chaotic dynamical systems, *J. Comput.*  
1052 *Phys.* **235**, 1 (2013).
- 1053 [106] N. Chandramoorthy, Z.-N. Wang, Q. Wang, and P. Tucker, Toward computing sensitivities of average  
1054 quantities in turbulent flows, *arXiv preprint arXiv:1902.11112* (2019).
- 1055 [107] T. Barth, On the role of error and uncertainty in the numerical simulation of complex fluid flows,  
1056 presented at the 2010 SIAM Annual Meeting, SIAM, Philadelphia (2010).
- 1057 [108] M. Nazarov and J. Hoffman, On the stability of the dual problem for high Reynolds number flow  
1058 past a circular cylinder in two dimensions, *SIAM J. Sci. Comput.* **34**, 1905 (2012).
- 1059 [109] Q. Wang and J.-H. Gao, The drag-adjoint field of a circular cylinder wake at reynolds numbers 20,  
1060 100 and 500, *J. Fluid Mech.* **730**, 145 (2013).
- 1061 [110] J. Hoffman, Computation of mean drag for bluff body problems using adaptive DNS/LES, *SIAM J.*  
1062 *Sci. Comput.* **27**, 184 (2005).
- 1063 [111] J. Hoffman, Adaptive simulation of the turbulent flow past a sphere, *J. Fluid Mech.* **568**, 77 (2006).
- 1064 [112] N. Jansson, J. Hoffman, and M. Nazarov, Adaptive simulation of turbulent flow past a full car  
1065 model, *Proceedings of the SC '11, ACM International Conference for High Performance Computing,*  
1066 *Networking, Storage and Analysis* , 20:1 (2011).