14th International Symposium on Turbulence and Shear Flow Phenomena (TSFP14)
Heidelberg, Germany, July 28–31, 2026

REINFORCEMENT LEARNING BASED CONTROL OF THE KELVIN HELMHOLTZ MODE IN 2D INCOMPRESSIBLE MAGNETOHYDRODYNAMICS

Sergio A. Angelini
Department of Mechanical and Aerospace Engineering
The University of Manchester
Oxford Rd, Manchester M13 9PL, UK
sergio.angelini@postgrad.manchester.ac.uk

Alex Skillen
Department of Mechanical and Aerospace Engineering
The University of Manchester
Oxford Rd, Manchester M13 9PL, UK
alex.skillen@manchester.ac.uk

Małgorzata J. Zimoń
IBM Research Europe
IBM
IBM Research Europe, Daresbury WA4 4AD, UK
malgorzata.zimon@uk.ibm.com

Mingfei Sun
Department of Computer Science
The University of Manchester
Kilburn Building, Oxford Rd, Manchester M13 9 PL, UK
mingfei.sun@manchester.ac.uk

Wei Pan
Department of Computer Science
The University of Manchester
Kilburn Building, Oxford Rd, Manchester M13 9 PL, UK
wei.pan@manchester.ac.uk

ABSTRACT

We investigate the active stabilisation of the Kelvin-Helmholtz (KH) instability in a two-dimensional, incompressible, inductionless magnetohydrodynamic (MHD) shear layer at $Re = 1000$ using Deep Reinforcement Learning (DRL). While externally applied magnetic fields are known to suppress shear-driven instabilities via electromagnetic damping, classical results from linear stability analysis prescribe stability thresholds based on constant actuation. In this work, a Proximal Policy Optimisation (PPO) agent is trained to minimise perturbation energy while penalising excessive control effort, utilising a set of localised vorticity probes and flow statistics as state observations. Numerical simulations are performed using the `BOUT++` framework at $Re = 1000$, with the linear phase of the instability compared against Orr–Sommerfeld eigenvalue analysis. To characterise the learnt policy, we perform a series of statistical analyses, identifying a clear hierarchy in adaptability between the RL agent, a Bayesian-optimised Proportional–Integral–Derivative (PID) controller, and a constant baseline. Our results show that the RL controller achieves stabilisation with low actuation amplitude through a policy possessing internal temporal structure, with late stage dynamics well described by autoregressive behaviour. At the same time, the reduced observable examined here provides only weak unique predictive information beyond the recent control history, indicating that the learnt strategy cannot be readily explained through feedback based solely on the reduced state proxy. In contrast, the optimised PID controller converges to a largely uniform response that does not account for variability in the initial perturbation. These findings demonstrate that reinforcement learning can discover effective control laws for MHD shear flows that extend beyond constant forcing and classical feedback strategies.

INTRODUCTION

The Kelvin-Helmholtz (KH) instability is a fundamental mechanism driving the transition of shear layers to turbulence in a wide range of fluid systems, including atmospheric flows (Luce *et al.*, 2010), astrophysical plasmas (Masson & Nykyri, 2018), and liquid-metal applications (Sharma *et al.*, 2023). In magnetohydrodynamic (MHD) settings, externally applied magnetic fields can suppress the growth of KH modes through electromagnetic damping (Chandrasekhar, 1961), (Keppens *et al.*, 1999), (Miura & Pritchett, 1982), providing a potential route for active flow control. While clas-

sical approaches rely on fixed-form feedback laws, such as Proportional-Integral-Derivative (PID) controllers, recent advances in reinforcement learning (RL) applied to fluid dynamics (Kim *et al.*, 2024) suggest the possibility of discovering control strategies directly from data. In this direction, the present work expands the scope of applicability of RL methods to include the active stabilisation of the Kelvin-Helmholtz instability in an MHD shear layer. Moreover, we evaluate not only the efficiency of RL-based stabilisation against classical baselines but also the underlying structure of the strategies that emerge. Through statistical analysis of the control trajectories, we show that the learned policy develops an internal temporal organisation and achieves effective stabilisation with low actuation amplitude, while not being reducible to either constant forcing or scalar feedback based on the reduced observable considered here. By contrast, the optimised classical baselines converge to comparatively uniform responses, reflecting the limitations imposed by control laws of fixed functional form.

MODEL FORMULATION

We consider the motion of an incompressible, electrically conductive fluid in a two dimensional domain Ω, periodic along the z direction and wall bounded along x. The dynamics are characterised by the magnetic Reynolds number $Re_m = \frac{U\delta}{\eta}$, which measures the ratio of magnetic advection by the flow to magnetic diffusion by resistivity, where U and δ denote characteristic velocity and length scales, and η is the magnetic diffusivity. In the asymptotic regime $Re_m \ll 1$, magnetic diffusion dominates induction, so the induced magnetic field adjusts quasi instantaneously to the velocity field and the full induction equation reduces to an elliptic constraint. As the present analysis is restricted to this inductionless limit, the governing MHD equations may be written as:

$$\begin{aligned} \partial_t \omega + (u\cdot\nabla)\,\omega &= \frac{Ha^2}{Re}\nabla\times(J\times B) + \frac{1}{Re}\nabla^2\omega, \\ J &= u\times B - \nabla\varphi, \\ \nabla\cdot B &= 0, \\ \nabla\cdot u &= 0, \\ \nabla^2\varphi &= \nabla\cdot(u\times B), \end{aligned} \tag{1}$$

where ω is the vorticity, B the externally applied magnetic field, J the induced current density, φ the electric potential, ∇^2 the Laplacian, $\nabla\times$ the curl, $\nabla\cdot$ the divergence, and ∇ the gradient operator. Each variable has been nondimensionalised with respect to the shear layer width δ of a reference profile $V = \tanh(x/\delta)$, half the velocity jump $U = |V_{-\infty} - V_{\infty}|/2$ (with $V_{\pm\infty}$ being the velocity of the flow far from the shear layer), and a reference magnetic field strength B_0. The relevant parameters are the Reynolds number $Re = U\delta/\nu$ and he square of the Hartmann number $Ha^2 = \sigma B_0^2\delta^2/\rho\nu$. In what follows, we refer to the ratio $N = Ha^2/Re$ as the interaction parameter (also known as the Stuart number), which is expressed in terms of the instantaneous B_0 as a fraction of the maximum allowed $B^* = \max B$, since the control action is not constant in general. The variables ν, ρ, and σ denote, respectively, the kinematic viscosity, density, and electrical conductivity of the fluid.

Perfect-conductor boundary conditions are imposed at the walls in x, so that $\varphi|_{\partial\Omega} = 0$, together with free-slip conditions for the velocity field. In the two-dimensional setting considered here, the right-hand side of the Poisson problem for φ vanishes identically, and an application of the Maximum Principle then yields $\varphi = 0$ throughout Ω, further simplifying the system. The magnetic field B is obtained from the vector potential A, defined via $B = \nabla\times A$, generated by a pair of coaxial coils in a Helmholtz configuration. This formulation has the advantage that the solenoidal condition on B is satisfied automatically, provided A is known. As the present work is concerned with the field generated near the centre of the coil pair, we idealise the configuration as two coaxial rings of current density j_i, $i = 1,2$, and the associated field is computed through numerical integration of the corresponding elliptic integrals.

NUMERICAL SIMULATIONS

We implement the system of equations defined in Eq. (1) using the plasma simulation framework `BOUT++` (Dudson *et al.*, 2020), fixing the Reynolds number $Re = 1000$. The governing equations are solved on a two-dimensional rectangular domain defined by the Cartesian product $x \in [0,1],\ z \in [0,1]$, where the z-direction is treated as periodic, while the x-direction is bounded. This geometry is chosen to capture the essential features of the underlying dynamics while maintaining computational efficiency.

Spatial discretisation is performed using a uniform grid with $N_x = 128$ points in the bounded direction and $N_z = 64$ points in the periodic direction, corresponding to grid spacings $\Delta x = L_x/N_x$ and $\Delta z = L_z/N_z$. Along the x-direction, first- and second-order derivatives are approximated using second-order central finite difference schemes, while upwind, third order methods are employed for advective terms. In the periodic z-direction, spatial derivatives are computed using Fourier spectral methods based on Fast Fourier Transforms (FFTs), providing high accuracy for smooth solutions. Time integration is carried out using an explicit fourth-order Runge–Kutta (RK4) scheme. Concerning the external magnetic field, we consider a pair of coils located at

$$(R_{\text{coil}}, Z_{\text{coil},1}) = (2.5,\ 1.75), \qquad (R_{\text{coil}}, Z_{\text{coil},2}) = (2.5,\ -0.75),$$

with a current chosen to produce a central vertical magnetic field of $B_{\text{centre}} = 1$ T. The corresponding coil current is computed self-consistently using the vacuum permeability μ_0, and the magnetic flux function is evaluated using Green's functions. The equilibrium structure of the flow is prescribed as an equilibrium shear layer $V = \tanh(x/\delta)$.

The unstable modes of the selected profile are found through eigenvalue analysis of the relevant Orr-Sommerfeld equation, as conveyed in Figure 1 for the profiles of V and B. These results, obtained through numerical simulations, are in accordance with established literature, conveying how the magnetic field acts as a damping force to suppress the mode's growth after a threshold is violated. Setting $\delta = 0.05$ and recalling that the periodic domain has length $L_z = 1$, the fundamental Fourier mode admitted by the domain corresponds to a wavenumber $k = 2\pi/L_z$, giving $k\delta \approx 0.314$. This value lies within the unstable band identified for $N = 0$, which extends up to $k\delta \approx 0.9$ (see the right panel of Figure 1), and it is therefore the natural choice for the perturbation wavenumber. The initial perturbation ω' is taken as a Gaussian profile localised at the shear layer, modulated by a sinusoid at the fundamental wavenumber,

$$\omega' \propto \exp\left(-\frac{(x-x_0)^2}{2\delta^2}\right)\sin(kz).$$

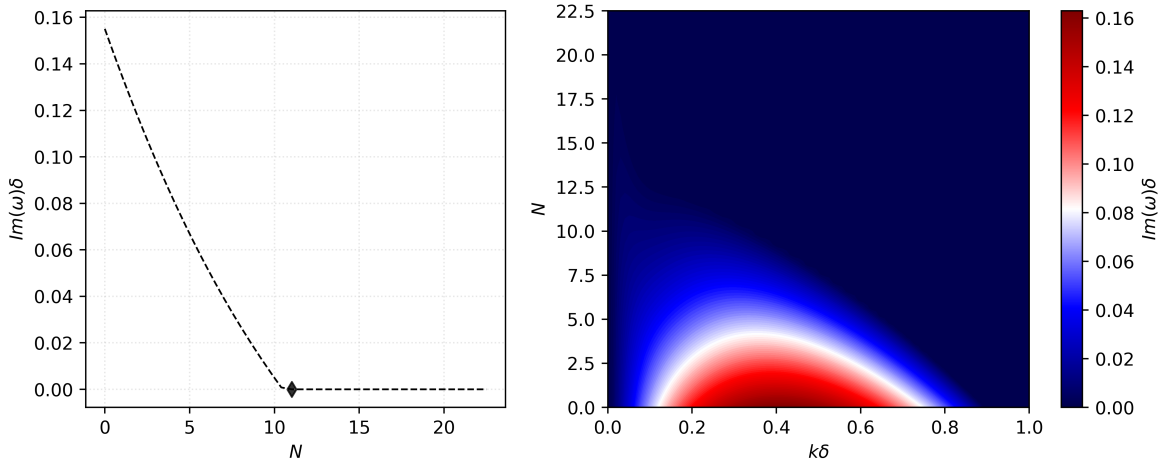

Figure 1. Numerical results of the Orr-Sommerfeld problem computed using the DEDALUS framework. *On the left*: Growth rates curve for the mode $k\delta \approx 0.314$, with the growth rate vanishing at the critical interaction parameter, denoted by the black diamond. *On the right*: Stability analysis of the profile for $Re = 1000$. The stabilising effect of increased values of N is evident as the growth rates converge to 0.

It is important to note that this initialisation does not correspond to the most unstable eigenmode associated with $k\delta \approx 0.314$; as a consequence, the measured growth rate from the simulations should not be expected to match the theoretical linear prediction exactly. Because the perturbation is comprised of a mixture of eigenmodes rather than a pure unstable eigenfunction, the linear theory result for the dominant eigenmode acts as an upper bound on the growth rate observed in the nonlinear simulations. This is confirmed by a projection of the initialisation onto the Orr Sommerfeld eigenbasis, which demonstrates that approximately 80 eigenmodes are required to reconstruct the Gaussian perturbation to machine precision (where the reconstruction error is defined as the maximum of the absolute error between the reconstructed and the original function $\varepsilon \simeq \mathcal{O}(10^{-15})$), indicating that the initial condition has significant support across both unstable and stable modes of the spectrum and cannot be well approximated by any small subset of eigenfunctions. This expectation is confirmed by a convergence study performed on three successively refined grids: $(N_x, N_z) \in \{128 \times 64,\ 256 \times 128,\ 512 \times 256\}$. Across all three grids, the simulations yield a consistent growth rate $\gamma = \mathrm{Im}(\omega)\,\delta \approx 0.123$, with a relative variation of approximately 2% between the coarsest and finest resolutions. This value is roughly 15% smaller than the growth rate associated with the most unstable eigenmode at $k\delta \approx 0.314$, consistent with the expectation that the theoretical bound should serve as an upper bound for the mixed-mode initialisation. We therefore conclude that the baseline discretisation employed to train the agent (the 128×64 grid) is sufficient to faithfully capture the linear phase of the instability's growth, which is precisely the dynamical regime the agent is designed to control. Finally, we perform a parameter scan in N to identify the threshold value above which the (ideal) linear growth rates vanish, as shown in the left panel of Figure 1. We note that this threshold constitutes a conservative upper bound on the stabilisation requirement; since the nonlinear simulations exhibit a growth rate already suppressed relative to the ideal eigenvalue, smaller values of N are expected to be sufficient to arrest the instability in practice. Based on this analysis, we select $N_{max} = 12$ as the interaction parameter for the training environment. This value, therefore, provides the agent with sufficient margin to discover non trivial stabilisation strategies without reducing the control problem to a trivial one.

REINFORCEMENT LEARNING

Reinforcement Learning formalises the idea of learning effective control strategies for dynamical systems through repeated interaction (Sutton & Barto, 2015). In this framework, a controller, generally known as *the agent*, interacts repeatedly with a dynamical system, referred to as *the environment*, to minimise some cost function associated with the latter. The interactions between the agent and the environment take place at a discrete set of points in time: as the dynamical system evolves, the agent receives an observable of the system, known as the *state*, and selects an action from the set of those available at that instant, which may itself be a random variable. Once selected, the action is fed to the environment, which advances the dynamical system in time and returns a *reward* signal used to update the agent. The agent therefore realises either a deterministic mapping or a stochastic one from the set of possible states $\mathscr{S}$ to the set of possible actions $\mathscr{A}$. In either case, we refer to this mapping as the *policy*, and the problem of finding an optimal controller is naturally cast in the framework of statistical learning. Recent years have seen the development of a multitude of RL methods based on Deep Neural Networks (DNNs), which have proven capable of reaching superhuman performance in a multitude of tasks including Chess and Go (Silver *et al.*, 2017). Given that the policy is represented by a DNN, the central challenge becomes how to update its parameters effectively from the stream of rewards received during interaction, so as to take full advantage of the approximation capabilities of these networks. In this direction, Proximal Policy Optimization (PPO) (Schulman *et al.*, 2017) has emerged as a a widely used algorithm for continuous, possibly multi dimensional action spaces. PPO is a policy gradient method designed to implement the largest possible improvement step on the policy without causing its performance to collapse. It achieves this by enforcing proximity between the updated policy and the previous one through a clipped objective function. If the probability of selecting a given action under the new policy becomes larger than under the old one (beyond a speci-

fied threshold), the incentive to increase it further is capped, ensuring that updates remain conservative and that the destructive policy shifts that destabilise training in other methods are avoided. In this work, we define and train a Multi-Layer Perceptron (MLP) architecture using the pipeline provided by `stable-baselines3` (Raffin *et al.*, 2021) with the aim of learning a state dependent control of the underlying physical system.

CONTROL PROBLEM

We study the suppression of the Kelvin-Helmholtz mode as an obstacle avoidance problem. To this end, we define a set of 96 probes, equally spaced in the z direction and restricted to the interval $[0.4, 0.6]$ in x, whose vorticity readings define the state of the system. These readings are concatenated with an estimate of the growth rate γ and of the L^1 norm of the perturbation's velocity as computed at the probes' locations, and fed at each time step to the PPO agent, which specifies a continuous action $a_t \in [0,1]$ corresponding to the relative intensity of the magnetic field B generated by the coils. Since the agent provides control values at discrete instants while the system evolves continuously, the control actions are interpolated using piecewise linear functions. We note that although the time derivative of the resulting field B is not guaranteed to be continuous, this does not constitute an issue: in the inductionless approximation, B is effectively decoupled from the fluid motion, and the system responds only to the instantaneous value of the field rather than its rate of change. Once the updated field intensity is received, the simulation advances by one time step and awaits the next action, and the cycle repeats until the end of the episode. For the problem considered in this work, the agent is trained to maximise the following reward:

$$r_t = \begin{cases} -\alpha \int_{\mathscr{P}} |u_x| - \beta a_t^2 - \mu (a_t - a_{t-1})^2 & \text{if } \int_{\mathscr{P}} |u_x| \, \mathrm{d}x < l, \\ -k & \text{if } \int_{\mathscr{P}} |u_x| \, \mathrm{d}x \geq l, \end{cases}$$

where u_x is the transverse velocity component, α, β, μ, k and l are tunable constants, and the integral is evaluated at the probe locations. The threshold l acts as a hard boundary on the agent's performance: should the monitored quantity exceed it at any point during an episode, the system is deemed to have left the admissible region, and the agent incurs a penalty $-k$. The value of l therefore concretises our definition of marginal stability. Smaller values of l encourage the agent to intervene early and aggressively at the first sign of perturbation growth, whereas larger values allow for more permissive policies. Whenever the system remains within the admissible region, the agent accrues a reward that is negatively proportional to the integrated measure of the estimated transversal velocity, combined with a further penalisation term, which discourages excessive control and discontinuous behaviour. This specification is directly motivated by the stability analysis of the preceding sections: without a penalty on the control magnitude, the agent would collapse onto a constant action, since the system has been engineered to be linearly stable at $a_t = 1$ (corresponding to $N = N_{max} = 12$). The requirement to minimise instantaneous power consumption instead drives the agent towards a family of non-trivial stabilisation strategies that exploit the available margin identified in the linear analysis.

Concerning training, we instantiate $K = 10$ parallel environments, each running an independent simulation. To prevent overfitting to a single trajectory and encourage the emergence of robust policies, the perturbation amplitude $\varepsilon \in [0.001, 0.01]$, the shear layer location $x_{\text{shear}} \in [0.4, 0.6]$, and the perturbation's phase $\phi \in [0, \pi]$ are randomised at the beginning of each new episode. For the runs discussed in the following Section, we have chosen $l = 0.02$, $\beta = 0.01$, $\mu = 0.01$, $\alpha = 1.0$, and $k = 3.0$. The MLP is trained until no appreciable variation in its mean reward is obtained, which we find corresponds to mean episode lengths that are coherent with the maximum prescribed duration of the simulations.

RESULTS

In this section, we analyse the policy learnt by the PPO agent through a series of statistical tests designed to assess the extent to which the control actions depend on the system state. To this end, we perform $M = 28$ simulations across varying grid resolutions, with $N_x \in \{128, 256, 512\}$ and $N_z \in \{64, 128, 256\}$, in order to evaluate the robustness of the learnt policy with respect to spatial discretisation. As a baseline, we tune a PID controller whose gains are optimised via Bayesian hyperparameter search using Gaussian process regression in the framework provided by `scikit-learn` (Pedregosa *et al.*, 2011). We note that the strategy learnt by the RL agent is stable across the set of grids considered. We then compare three control strategies: constant actuation, optimised PID control, and the learnt RL policy. The analysis focuses on establishing a clear hierarchy between the three controllers when identifying how the strategies utilise an informative proxy of the global state of the system. In particular, we show that the advantage of the RL agent stems from its ability to enable state dependent control adjustments beyond those captured by state agnostic controllers.

Learnt control strategy

To characterise the control strategy learnt by the PPO agent, we analyse the relationship between the control action and the system state across multiple trajectories. In particular, we aim to determine whether the policy exhibits genuine state dependent behaviour, as opposed to following a predominantly open-loop or pre-scheduled control sequence. The left panel of Figure 2 conveys the distribution of control trajectories for the RL model. We begin by noting how the maximum control $N(t) \approx 6$ exhibited by the agent lies well below the maximum available actuation $N_{max} = 12$ that was chosen based on linear stability theory considerations. Indeed, a control authority of $N_{max} = 12$ was expected to be a generous upper bound on the actual intensity required to achieve stabilisation. We find this initial evidence encouraging, and therefore proceed with a statistical analysis of the control trajectories. In this direction, we begin by isolating the L^1 norm of the perturbation's velocity at time t, hereby denoted as $||u_x(t)||_1 = \int_{\mathscr{P}} |u_x(t)| d\omega$ as a reduced state observable. We note that any subsequent result should be interpreted as a *lower bound* on the dependency of the policy on the state, as the selected proxy is but a portion of the effective state available to the agent. A first inspection based on the direct correlation between the control amplitude a_t and $||u_x(t)||_1$, suggests a moderate statistical dependence. We compute the pooled within trajectory correlation between the control amplitude and the state proxy $||u_x(t)||_1$ over the first 10% of the control steps, after removing the mean of each trajectory. We find a moderate positive correlation ($r \approx 0.51$), indicating that stronger than average actuation is consistent with larger perturbations in the early phase of the dynamics.

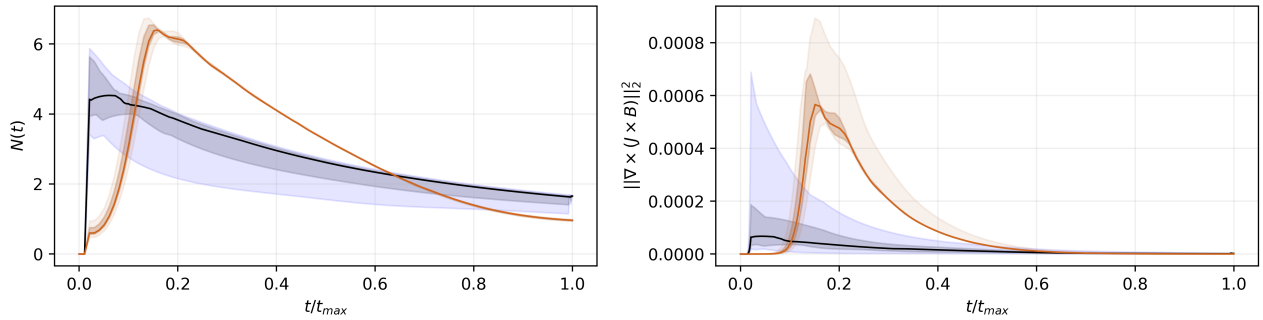

Figure 2. *On the left*: Distribution of the control amplitude for $M = 28$ runs. The solid black/orange lines represents the median control for the RL agent/PID controller, whereas the shaded region conveys the 90% confidence interval for each. *On the right* L^2 norm of the curl of the Lorentz force in the unfolding of the simulations. The correlation reflects the direct dependence of magnetic forcing on the applied field intensity.

However, correlation alone is not a definitive proof of genuine feedback. In particular, it does not distinguish causal state-responsive updates from common transient evolution or temporal autocorrelation. Moreover, over the full trajectory both quantities undergo a monotonic decay as the instability is suppressed, which can generate spurious associations. To account for this, we analyse detrended signals and find that the apparent correlation becomes statistically insignificant, indicating that the raw dependence is primarily driven by shared temporal relaxation rather than direct feedback. With the aim of understanding whether less obvious feedback mechanisms exist, we instead consider the increments of the control signal, $\Delta a_t = a_t - a_{t-1}$, which capture how the policy adjusts its action over time, fitting a kernel ridge regression to account for nonlinearity. In this formulation, including the state proxy marginally improves predictive performance, demonstrating that the system state correlates with how the control is updated. While the overall effect is modest when averaged over entire trajectories, the time-resolved analysis depicted in Figure 3 reveals a stronger dependence during the early stages of the evolution. However, this interpretation changes once the recent history of the control signal is explicitly accounted for. To this end, we perform a transfer entropy test (Schreiber, 2000), comparing predictions of future control increments based on past control increments alone with predictions augmented by the state proxy. The resulting gain in predictive accuracy is negligible, indicating that the reduced proxy $||u_x(t)||_1$ does not contribute unique information beyond what is already contained in the control history. This behaviour is consistent with the physical mechanism of instability stabilisation at work in the physical system described by the simulations: the streamwise component of the forcing magnetic field acts as a damping term directly proportional to the velocity of the transverse perturbation. Therefore, as the state proxy's magnitude decreases over time due to the actuation of the control, the variability of the environment diminishes, and different control trajectories coalesce, giving rise to the observed autoregressive behaviour. During the initial phase of the control problem, the larger variability of the uncontrolled dynamics gives rise to stronger statistical associations between the control signal and the reduced observable. However, the conditional forecasting analysis indicates that this coarse proxy contributes little unique predictive information once the recent control history is taken into account. Overall, these results show that the PPO agent has learnt a stabilising strategy with a pronounced internal temporal structure, rather than relying on aggressive actuation. At

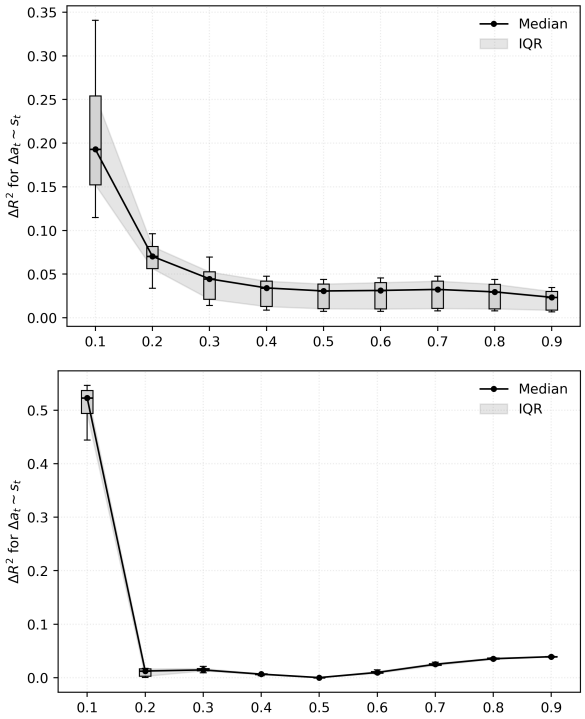

Figure 3. Comparison between the R^2 scores of the control signal's increments in time. *On the top*: MLP agent, the correlation between the state proxy and the control delta decreases in time, showcasing high initial variability. *On the bottom*: PID controller, the correlation decays immediately after the initial timesteps of the control runs considered, showcasing limited persistence of state-dependent feedback beyond the initial transient.

the same time, the weak dependence observed with respect to the chosen reduced observable suggests that the relevant decision variables may lie in higher-dimensional features of the flow not captured by the present diagnostic.

Comparison with PID and constant action

To contextualise the behaviour of the agent, we compare it against two baseline strategies: a constant actuation policy and an optimally tuned PID controller. The constant policy provides a reference for purely open loop control, while the PID

controller represents a classical feedback strategy based on a reduced scalar observable of the system state. The PID gains are tuned via Bayesian optimisation to minimise a cost function combining tracking error, defined as the instantaneous $||u_x(t)||_1$, control effort, and penalties for unstable trajectories. Despite this optimisation, the PID controller operates on a fixed functional form, with control actions determined solely by the instantaneous value of the error signal and its temporal derivatives. We perform this comparison with the goal of showcasing a clear hierarchy in state adaptability between the three strategies. The constant policy clearly admits no dependence on the evolving system state, while the PID controller exhibits a short lived state dependence, primarily during the initial transient. Notably, the PID controller demonstrates negligible variance across different trajectories. The tuned gains effectively enforce a singular, environmentally invariant strategy that does not account for the parametric variations of the environment. In contrast, the RL policy exhibits a temporally structured form of state dependence, not constrained by a predetermined feedback structure. At the same time, the analysis provided shows that the proxy $||u_x(t)||_1$ contributes only weak predictive information beyond the intrinsic control history. This suggests that the learned policy does not rely strongly on the same reduced feature exploited by the PID controller, and instead either follows a robust internally structured strategy or responds to higher-dimensional aspects of the flow inaccessible to low-order baselines. Overall, the comparison shows that the PPO agent achieves stabilisation with low actuation amplitude through a control law that cannot be reduced to either constant forcing or scalar feedback based on the reduced state proxy $||u_x(t)||_1$.

CONCLUSIONS

In this work, we have investigated the stabilisation of the Kelvin-Helmholtz instability in two dimensional incompressible magnetohydrodynamics using reinforcement learning. Beyond demonstrating effective control, our primary focus has been on understanding the structure of the learnt policy and its dependence on the system state. The resulting analysis shows that the control signal possesses an internal temporal organisation, with late stage dynamics well described by autoregressive behaviour. At the same time, however, the specific reduced observable considered here $||u_x(t)||_1$ provides weak predictive information beyond the recent control history, indicating that the learnt strategy is not readily explained through this specific state proxy. A comparison with a Bayesian-optimised PID controller further clarifies this behaviour. While the PID exhibits strong initial feedback through the reduced observable, the negligible variance displayed across different trajectories is consistent with a largely uniform response imposed by its fixed functional form. By contrast, the reinforcement learning agent learns a more nuanced control structure, defining a control law that cannot be reduced to either constant forcing or scalar feedback based on the reduced state. This fact suggests that the relevant decision variables may lie in higher-dimensional flow features, or in temporal structures inaccessible to classical low-order controllers. Future work will assess the effect of reduced state observability and explore recurrent architectures, which offer a principled route to encoding hidden mode dynamics directly from temporal probe sequences.

REFERENCES

Chandrasekhar, Subrahmanyan 1961 *Hydrodynamic and hydromagnetic stability*.

Dudson, Benjamin Daniel, Hill, Peter Alec, Dickinson, David, Parker, Joseph, Dempsey, Adam, Allen, Andrew, Bokshi, Arka, Shanahan, Brendan, Friedman, Brett, Ma, Chenhao, Schwörer, David, Meyerson, Dmitry, Grinaker, Eric, Breyiannia, George, Muhammed, Hasan, Seto, Haruki, Zhang, Hong, Joseph, Ilon, Leddy, Jarrod, Brown, Jed, Madsen, Jens, Omotani, John, Sauppe, Joshua, Savage, Kevin, Wang, Licheng, Easy, Luke, Estarellas, Marta, Thomas, Matt, Umansky, Maxim, Løiten, Michael, Kim, Minwoo, Leconte, M, Walkden, Nicholas, Izacard, Olivier, Xi, Pengwei, Naylor, Peter, Riva, Fabio, Tiwari, Sanat, Farley, Sean, Myers, Simon, Xia, Tianyang, Rhee, Tongnyeol, Liu, Xiang, Xu, Xueqiao & Wang, Zhanhui 2020 BOUT++.

Keppens, R., Toth, G., Westermann, R. H. J. & Goedbloed, J. P. 1999 Growth and saturation of the kelvin-helmholtz instability with parallel and anti-parallel magnetic fields **61** (1), 1–19.

Kim, Innyoung, Jeon, Youngmin, Chae, Jonghyun & You, Donghyun 2024 Deep reinforcement learning for fluid mechanics: Control, optimization, and automation. *Fluids* **9** (9).

Luce, Hubert, Mega, Tomoaki, Yamamoto, Masayuki K., Yamamoto, Mamoru, Hashiguchi, Hiroyuki, Fukao, Shoichiro, Nishi, Noriyuki, Tajiri, Takuya & Nakazato, Masahisa 2010 Observations of kelvin-helmholtz instability at a cloud base with the middle and upper atmosphere (mu) and weather radars. *Journal of Geophysical Research: Atmospheres* **115** (D19).

Masson, A. & Nykyri, K. 2018 Kelvin–helmholtz instability: Lessons learned and ways forward **214** (4), 71.

Miura, Akira & Pritchett, P. L. 1982 Nonlocal stability analysis of the mhd kelvin-helmholtz instability in a compressible plasma. *Journal of Geophysical Research: Space Physics* **87** (A9), 7431–7444.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, E. 2011 Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830.

Raffin, Antonin, Hill, Ashley, Gleave, Adam, Kanervisto, Anssi, Ernestus, Maximilian & Dormann, Noah 2021 Stable-baselines3: Reliable reinforcement learning implementations. *Journal of Machine Learning Research* **22** (268), 1–8.

Schreiber, Thomas 2000 Measuring information transfer. *Phys. Rev. Lett.* **85**, 461–464.

Schulman, John, Wolski, Filip, Dhariwal, Prafulla, Radford, Alec & Klimov, Oleg 2017 Proximal policy optimization algorithms.

Sharma, Shubham, Chandra, Navin Kumar, Kumar, Aloke & Basu, Saptarshi 2023 Shock-induced atomisation of a liquid metal droplet. *Journal of Fluid Mechanics* **972**, A7.

Silver, David, Hubert, Thomas, Schrittwieser, Julian, Antonoglou, Ioannis, Lai, Matthew, Guez, Arthur, Lanctot, Marc, Sifre, Laurent, Kumaran, Dharshan, Graepel, Thore, Lillicrap, Timothy, Simonyan, Karen & Hassabis, Demis 2017 Mastering chess and shogi by self-play with a general reinforcement learning algorithm.

Sutton, R.S. & Barto, A.G. 2015 *Reinforcement Learning: An Introduction*. MIT Press.