Bio-inspired Wake Tracking for Aircraft Formation Flight Based on Reinforcement Learning

Ignace Ransquin*, Philippe Chatelain†
Institute of Mechanics, Materials and Civil Engineering
Université catholique de Louvain, 1348, Louvain-la-Neuve, Belgium

Formation flight is known for improving the overall aerodynamic efficiency of a pair of aircraft. This work presents a bio-inspired tracking strategy that correctly positions the follower in the wake of the leader in order to optimize the benefits of formation flight, i.e. optimize apparent drag reduction. A simplified aerodynamic model based on Prandtl's lifting line theory enables to compute the 6 degrees-of-freedom dynamics of the follower under the influence of a leader wake. Those dynamics are controlled over time thanks to a speed-control autopilot, while a neural network modifies the speed target according to the follower dynamics. Indeed, the follower dynamic measurements are used to train the neural network through the reinforcement learning framework, drawing inspiration from the trial and error of animal learning. The use of the follower dynamics as unique control measurements ensures a total independence of the tracking method from direct measurements of the leader wake position. Simulations demonstrate the good performances of this new reinforcement learned tracking method when dealing with a simple two-aircraft formation as well as with more challenging configurations involving a larger flock.

Nomenclature

$\mathbf{a}_n$	=	action vector
$\mathcal{R}$	=	aspect ratio
b	=	wing span
Δt_{update}	=	learning time step
$\mathbf{F}_f = [F_x, F_y, F_z]^T$	=	aircraft forces
Γ	=	circulation
$\mathbf{M}_f = [M_x, M_y, M_z]^T$	=	aircraft torques
$(ox_i y_i z_i)$	=	inertial frame
$(ox_{ac} y_{ac} z_{ac})$	=	aircraft frame
$\mathbf{o}_n$	=	observation vector
ϕ, θ, ψ	=	angles of roll, pitch and yaw
p, q, r	=	roll, pitch and yaw rates
r_n	=	reward
ρ	=	air density
$\mathbf{u}_c = [\delta_{Ail}, \delta_{VTP}, \delta_{HTP}, \delta_{th}]^T$	=	control commands
U_∞	=	reference velocity
$\mathbf{V}_a = [U_a, V_a, W_a]^T$	=	relative air speed
$\mathbf{V}_f = [U_f, V_f, W_f]^T$	=	follower aircraft speed
$\Delta\dot{\mathbf{V}}_f = [\Delta\dot{U}_f, \Delta\dot{V}_f, \Delta\dot{W}_f]$	=	Acceleration differences in translation

*Ph.D. student; ignace.ransquin@uclouvain.be.
†Professor; philippe.chatelain@uclouvain.be.

$\Delta\dot{\omega}_f = [\Delta\dot{p}, \Delta\dot{q}, \Delta\dot{r}]$	=	Acceleration differences in rotation
$\mathbf{x}_f = [x_f, y_f, z_f]^T$	=	follower aircraft position
$[.]_i^T$	=	components in the inertial frame
$[.]_{ac}^T$	=	components in the aircraft frame

I. Introduction

Formation flight is known to improve the efficiency of a two-aircraft formation (leader+follower), as proven by early investigations [1, 2]. Indeed, the follower benefits from substantial apparent drag reduction when flying in the upwash region of the leader wake. This bio-inspired technique can therefore lead to significant fuel savings as showed either in a military context [3, 4] or more recently in commercial aviation [5].

Whereas birds usually fly in close formation, civilian applications rather target the concept of extended formation flight [6], that better adresses safety concerns. In that case, the follower is flying several tens of wingspans downstream of the leader. The follower still benefits from a positive interaction as the leader wake vortices are strong, long-lasting structures. However, simplified linear analysis [7] has shown that 50% of the benefit is lost if the follower cannot maintain a lateral and vertical relative position within 10% span of the leader, wings tip to tip. This has also been confirmed experimentally [8]. Because of that sensitivity to position, an accurate tracking of the wake position is thus mandatory in order to perform efficient extended formation flight.

The tracking of the wake vortices is not a new concept, yet most of the methods studied were so far focusing on the detection of wake hazards in the airspace or runways surroundings. The direct measurement of the flow field has been envisioned, through the use of an on-board LIDAR [9], or ground-based acoustic devices [10]. However, such complex technologies pose a significant challenge for on-board operational use also from the perspective of both cost and weight. Alternative solutions have been envisioned through the processing by an Ensemble Kalman Filter (EnKF) of distributed pressure measurements in order to determine parameters of the wake that influence the follower [11] (i.e. position and circulation). However, this technique also suffers from the need of specific instrumentation to be installed on board of the follower. This has led the authors of [12] to conduct a similar study based on measurements coming from simplified dynamics of the follower. That use of flight dynamics measurements showed promising results for the detection of the wake parameters [12] and motivated the development of a wake tracking strategy based on the 6 degrees-of-freedom (6DOF) dynamics of the follower * [13].

Drawing inspiration from a bird that learns formation flight through experience, the present work focuses on the tracking of the optimal position in the wake of a leader through the use of a deep learning algorithm within an adapted simulation framework. Indeed, all the methods described in [11–13] require the use of wake vortex models. Learning how to fly in the wake of one or even several leaders offers an adaptative solution able to react to unpredicted behaviors or complex environment without a predetermined model of the vortices. The use of deep learning algorithms for flow problems is not new, and covers wide applications from flow classification [14, 15] and forecasting of turbulent environment [16], to the load alleviation on wind turbines [17]. Among those algorithms, the Reinforcement Learning (RL) approach is used in [18] where fishes learn to school, leading to collective energy savings. The authors of [19] manage to teach a glider how to exploit thermal soaring with a similar approach. The RL framework gives therefore promising results in energy saving applications, but has never been applied to the challenging problem of wake tracking we tackle in this work.

The present study is intended to investigate the use of a RL technique to teach a follower how to fly optimally in the wake of a leader in order to benefit from the maximum induced lift. The learning process

*Confidential: Patent submited in July 2020 uses a model based on the Prandtl lifting line theory in order to characterize the aerodynamics of the follower. It also captures the leader wake vortices. The follower uses its translational and rotational accelerations measurements to determine the speed required to converge towards an optimal position. The designed strategy is first tested on a two-aircraft formation. We show that the follower is able to find the sweet spot in the wake of the leader. We then demonstrate the performance of the tracking strategy on a larger formation that follows the oscillating movement of a leader. We finally discuss the benefits and drawbacks of the proposed approach.

This paper is organized as follow. In Section II, we present the aerodynamic model characterizing the process, and we then describe the training process associated to this framework. The results are presented in Section III, first by validating the learned tracking strategy in the case of a single aircraft flying in the wake of leader, second by assessing the performance of the strategy on a larger formation. We close this work with conclusions and perspectives in Section IV.

II. Methodology

The proposed tracking strategy is based on the structuring of the follower control into a hierarchy of tasks, from high-level to low-level ones. It takes the form of a two-block architecture, as shown in Fig. 1. An autopilot computes the control surfaces (i.e. the aileron δ_{Ail}, the rudder δ_{VT} and the elevator δ_{HT} deflections) and the thrust command δ_{Th} of the follower in order to make it fly at a reference speed V_f^{ref}. This reference is provided by a neural network that is trained thanks to reinforcement learning (RL).

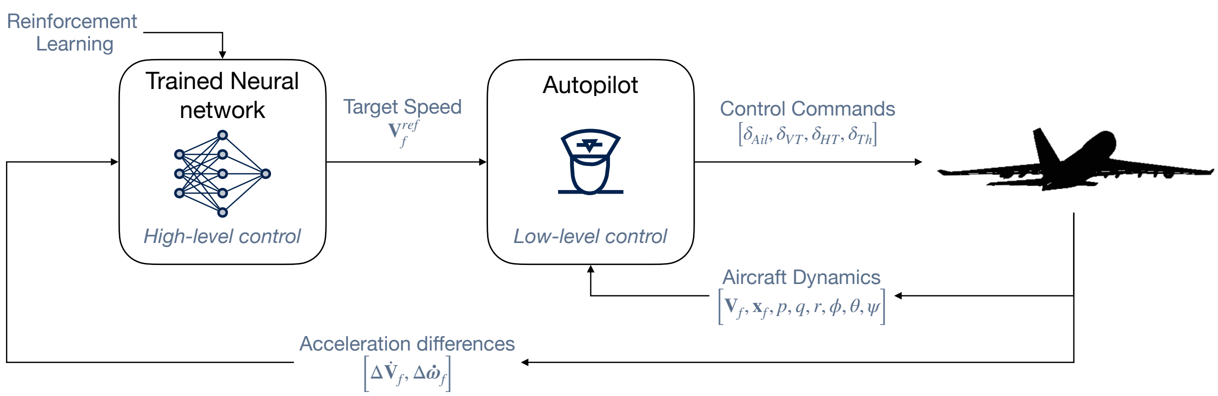


Figure 1 RL-based aircraft controller: the standard aircraft autopilot is fed with a target speed V_f^{ref} that guides a follower towards the optimal position in the wake of a leader. This target is provided by a high-level control made of a neural network trained thanks to reinforcement learning.

RL is a process by which the agent (in this case, a follower) learns to earn rewards through trial-and-error interactions with the environment (in this case, an atmosphere perturbed by a leader wake). This trial-and-error procedure requires to perform a large set of simulations of the interaction of the agent with its environment in order to get experience. While Large-Eddy Simulations (LES) of a two-aircraft formation, as presented in [12, 13], accurately models those interactions, they are not computationally affordable when given the number of samples required to converge the learning process. Therefore, we leverage some of the tools developed in [13], in particular a simplified and affordable 6DOF aerodynamic model of an aircraft which will be used for the learning procedure. First we describe this model, then we present the high-level control based on RL. We finally outline the learning process.

A. Aircraft aerodynamics and dynamics.

The Prandtl Lifting Line (PLL) theory constitutes the core of the aerodynamic model. As depicted in Fig. 2, the follower aircraft is located at the position $\mathbf{x}_f = [x_f, y_f, z_f]_i^T$ in the inertial frame $(ox_i y_i z_i)$ and comprises three straight lifting lines representing the wing, the vertical tail plane (VTP) and the horizontal tail plane (HTP). In this model, the wake of an aircraft is made of one flat vortex sheet produced by the wing, and the two perpendicular vortex sheets produced by the tail due to the HTP with the VTP. All those vortex sheets are assumed to go straight to infinity.

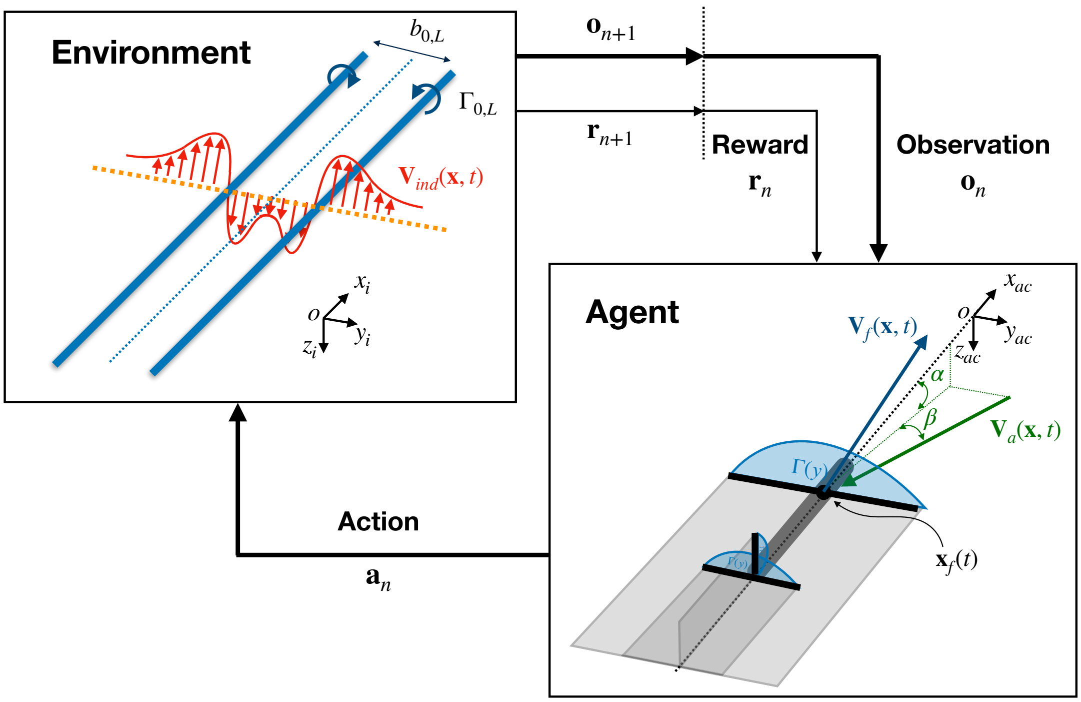


Figure 2 Reinforcement Learning (RL) framework made of the agent, the perturbed environment with its effect on the agent, and the RL iterations.

The wake of the leader is modeled as two counter-rotating vortex tubes of circulation $\Gamma_{0,L}$ and spacing $b_{0,L}$ inducing the upwash $\mathbf{V}_{\text{ind}}(\mathbf{x}, t)$ on the follower wing. Using the Prandtl lifting line theory, one computes the circulation distribution along each one of the lifting lines, solving the integro-differential equation that relates the circulation with the relative air velocities along the lines $\mathbf{V}_a(\mathbf{x}, t)$ (i.e. the combination of the velocities induced by the follower movement, the leader's wake and the lifting lines themselves). Note that this constitutes a quasi-steady assumption, where the wake instantaneously propagates to infinity. The forces $\mathbf{F}_f = [F_x, F_y, F_z]_i^T$ and the moments $\boldsymbol{M}_\text{f} = [M_x, M_y, M_z]_{ac}^T$ can be deduced thanks to the circulation distributions. The integration of the accelerations $\dot{\mathbf{V}}_f = [\dot{U}_f, \dot{V}_f, \dot{W}_f]_i^T$ (in the inertial frame $(ox_i y_i z_i)$) and angular accelerations $\dot{\boldsymbol{\omega}}_f = [\dot{p}, \dot{q}, \dot{r}]_{ac}^T$ (in the aircraft frame $(ox_{ac} y_{ac} z_{ac})$) allows the integration of the position, attitude and velocity over time [20]:

$$m\dot{\boldsymbol{V}}_f = \mathbf{F}_f \quad (1)$$

$$\dot{\boldsymbol{x}}_f = \mathbf{V}_f \quad (2)$$

$$\begin{cases} I_x\dot{p} - (I_y - I_z)qr = M_x \\ I_y\dot{q} + (I_x - I_z)pr = M_y \\ I_z\dot{r} - (I_x - I_y)pq = M_z \end{cases} \quad (3)$$

$$\begin{bmatrix} \dot{\phi} \\ \dot{\theta} \\ \dot{\psi} \end{bmatrix} = \begin{bmatrix} 1 & \sin\phi\tan\theta & \cos\phi\tan\theta \\ 0 & \cos\phi & -\sin\phi \\ 0 & \sin\phi\sec\theta & \cos\phi\sec\theta \end{bmatrix} \begin{bmatrix} p \\ q \\ r \end{bmatrix} \quad (4)$$

where $\mathbf{x}_f = [x_f, y_f, z_f]_i^T$ and $\mathbf{V}_f = [U_f, V_f, W_f]_i^T$ are the position and velocity of the aircraft, $[\phi, \theta, \psi]^T$ and

$\omega_f = [p, q, r]^T$ are the angular positions and angular body rates, m is the mass and I_x, I_y, I_z are the products of inertia.

We associate this complete 6DOF dynamics to the afformentioned autopilot designed to pilot the control commands of the aircraft (i.e. δ_{Ail}, δ_{VT}, δ_{HT} and δ_{Th}). Those control commands modify the aerodynamic loads on the lifting surfaces, which in turn modify the dynamics of the aircraft. The autopilot is designed to make the aircraft fly at a desired speed $\mathbf{V}_f^{ref} = [U_f^{ref}, V_f^{ref}, W_f^{ref}]_i^T$ in its environment; it is much akin to what is done in [21]. In the scope of this study, the aircraft always maintains a fixed longitudinal target speed (U_f^{ref} remains constant), while the targets for the vertical and lateral speeds (i.e. in the $(oy_i z_i)$ plane) change during the learning process, i.e. V_f^{ref} and W_f^{ref} vary over time. Indeed, as mentioned in Section I, the wake surfing benefit is highly sensitive to the lateral and vertical relative position, which justifies to investigate a learning strategy considering only the $(oy_i z_i)$ plane.

B. High-level control based on Reinforcement Learning

The high-level control of the wake tracking strategy consists in a deep neural network. It constantly modifies the target speed of the follower according to accelerations measurements. We use RL to train this neural network. The iterative procedure of the RL algorithm is presented in Fig. 2. At every learning step n, the agent gets the observation vector $\boldsymbol{o}_n$ from the environment and takes an action $\boldsymbol{a}_n$. The agent is given some reward r_{n+1} and gets a new observation vector $\boldsymbol{o}_{n+1}$. The ultimate goal of RL algorithms is to find the optimal mapping, called a policy $\pi^*(\boldsymbol{a}_n|\boldsymbol{o}_n)$, between the observations and the action, given an objective function $J(\pi)$. For each observation $\boldsymbol{o}_n$, the optimal policy maximizes the return $g_n = \sum_{k=0}^{\infty} \gamma^k r_{n+k+1}$, i.e. the sum of discounted future rewards, with γ the discount factor. Let us clearly define the observations, actions and reward as well as the RL algorithm used.

Observations, actions, reward

From the environment perspective, the agent is defined by the actions it performs. The follower is able to change its position within the perturbed environment, thanks to the autopilot and the modification of the target in the $(oy_i z_i)$ plane. The actions performed by the follower are therefore a modification of the targeted speeds V_f^{ref} and W_f^{ref} for the lateral and vertical behaviors respectively. This is the action space

$$\mathscr{A} : \boldsymbol{a}_n = \begin{cases} V_f^{ref} \\ W_f^{ref} \end{cases} . \tag{5}$$

Another fundamental aspect of the RL problem for wake tracking consists in the observations the follower can perform in its environment. This is the observation space $\mathscr{O}$. As we want to optimize the apparent drag reduction of the follower, compared to a similar aircraft that flies in the similar unperturbed conditions, we compute the accelerations $\dot{\mathbf{V}}_f$ and angular accelerations $\dot{\boldsymbol{\omega}}_f$ experienced by the follower when moving in its environment. Then, we compare them to the accelerations $\dot{\boldsymbol{V}}_f^{isol}$ and $\dot{\boldsymbol{\omega}}_f^{isol}$ of an identical but isolated aircraft, flying at the same speed, and subjected to the same control commands, but not subjected to the influence of a perturbed environment (i.e. flying in a clean atmosphere). This allows to compute the acceleration differences $\Delta\dot{\mathbf{V}}_f = \dot{\mathbf{V}}_f - \dot{\mathbf{V}}_f^{isol}$ and $\Delta\dot{\boldsymbol{\omega}}_f = \dot{\boldsymbol{\omega}}_f - \dot{\boldsymbol{\omega}}_f^{isol}$ which are the potential observations provided by the learning follower in its environment, as depicted in Fig. 3. Acceleration measurements are coherent with biological sensors as described in [19]. In addition, their variations when the follower moves in the $(oy_i z_i)$ plane are closely related to the proximity of wake vortices [22] which justifies their use in the context of wake tracking. In addition to the acceleration differences, the actual speeds of the aircraft are also used as observations. This gives all the potential observations

$$\mathcal{O} : \boldsymbol{o}_n = \begin{cases} V_f, W_f \\ \Delta\dot{U}_f, \Delta\dot{V}_f \\ \Delta\dot{p}, \Delta\dot{q}, \Delta\dot{r} \end{cases}. \tag{6}$$

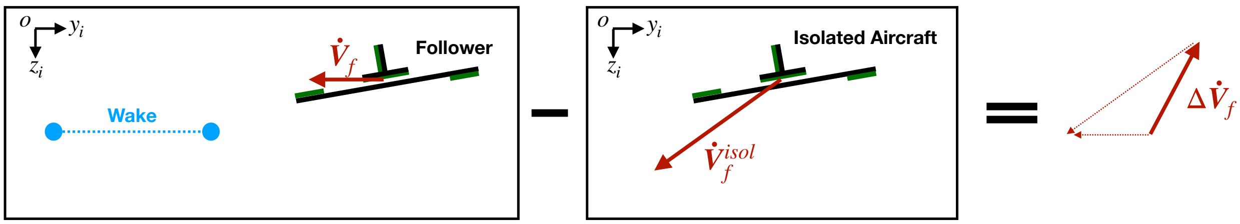


Figure 3 Compute of the difference of the accelerations perceived by the follower when flying in the wake of a leader.

We do not use the difference in vertical acceleration $\Delta\dot{W}_f$ as a potential observation, as we use it as the reward r_n provided by the environment to the agent. Indeed, the position of the maximum induced lift (and therefore maximal vertical acceleration) has been proven to be close to the optimal flying position in the wake of a leader [22]. Also, in order to fasten the convergence towards this optimal position, we get inspired by the work of Reddy [19] and we use a reward that linearly combines the difference in vertical acceleration $\Delta\dot{W}_f$ and the variation of this difference achieved over the subsequent time step. That is for better learning performance. We have

$$\mathcal{R} : r_n = \Delta\dot{W}_f + \tau\frac{d\Delta\dot{W}_f}{dt}, \tag{7}$$

where τ is a fine-tuned control parameter.

Algorithm

We use an off-policy model-free reinforcement learning algorithm called Soft Actor Critic (SAC) [23] in order to train the neural network. SAC is part of the maximum entropy reinforcement learning methods. It therefore uses an entropy augmented objective function

$$J(\pi) = \mathop{\mathrm{E}}_{\pi}\left[\sum_{n=0}^{\infty}\gamma^n r_n(\boldsymbol{o}_n, \boldsymbol{a}_n) - \alpha\log\left(\pi(\boldsymbol{a}_n \mid \boldsymbol{o}_n)\right)\right], \tag{8}$$

with α the positive entropy parameters. It explicitly controls the explore-exploit tradeoff, with higher α corresponding to more exploration, and lower α corresponding to more exploitation. SAC provides sample-efficiency learning and a low sensitivity to hyper-parameters while retaining the benefits of entropy maximisation and stability. This makes it a state-of-the-art model-free deep RL method [23].

We use TensorFlow [24] to build the fully connected neural network used in the tracking controller. It is made of 4 hidden layers with 64 neurons with relu as activation function. The size of the input layer depends on the number of observations used (this is discussed in Section III.A) while the output one comprises 2 neurons, one for each component of the action. The neural network is trained thanks to the open-source SAC implementation of Stable Baselines [25].

Learning process

The learning process implies going through two steps. The first one consists in training the aircraft so that it learns the optimal policy $\pi^*(\boldsymbol{a}_n|\boldsymbol{o}_n)$. For this first step, numerous learning episodes are performed using the aerodynamics model from Section II.A. At the beginning of each episode, the wake of the leader (assumed to always be the same) is reset at the same initial position in the (oy_iz_i) plane, and it starts to follow an oscillatory trajectory in that same plane, and whose amplitude and frequency is chosen randomly. Regarding the follower, it starts from a different position at each episode, and performs the actions asked by the learning algorithm. The performance of the follower will slowly increase as we go along the successive episodes. An episode ends when an observation $\mathbf{o}_n$ exceeds given boundaries, or when it has reached 200 learning steps (i.e. a time interval Δt_{update}).

The second step consists in testing the trained policy. This consists only in exploiting what has been learned. This is done by using once again the aerodynamic model, in order to compute the aerodynamic performances achieved by the follower (i.e. consumption) when it flies in the wake of a leader.

III. Results

In this section, we first focus on the result of the training of the neural network, starting with an analysis of the acceleration differences. We investigate their relevance in the context our RL-based tracking strategy. We then use the trained tracking strategy for a follower first flying in the wake of a fixed leader, then following the wake of a moving one. The performances of the controller are finally discussed. For all the subsequent simulations, we assume that all the aircraft are identical and fly at the cruise speed U_∞. Their geometrical parameters are given in Table 1.

	Wing	**HTP**	**VTP**
Span	b	$b_h/b = 0.26.$	$b_v/b = 0.22$
Aspect ratio	$Æ = 7.5$	$Æ_h = 4$	$Æ_v = 4$
Chord distribution	Elliptic	Rectangle	Rectangle
Incidence	$\alpha_{0,w} = 4°$	$\alpha_{0,h} = 1.42°$	$\alpha_{0,v} = 0°$
Lift Slope	$a_0 = 2\pi$	$a_0 = 2\pi$	$a_0 = 2\pi$
main wing lift factor	$C_{L,w}/Æ = 0.1$		

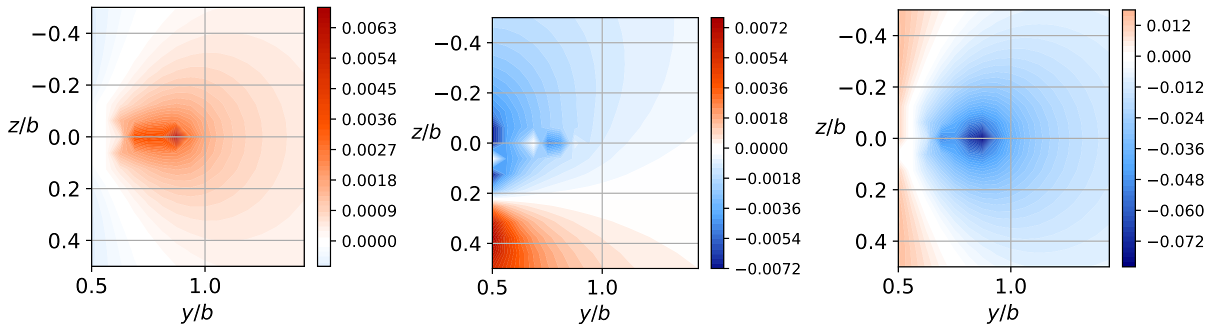
Table 1 Geometrical parameters of the aircraft

A. Training of the neural network

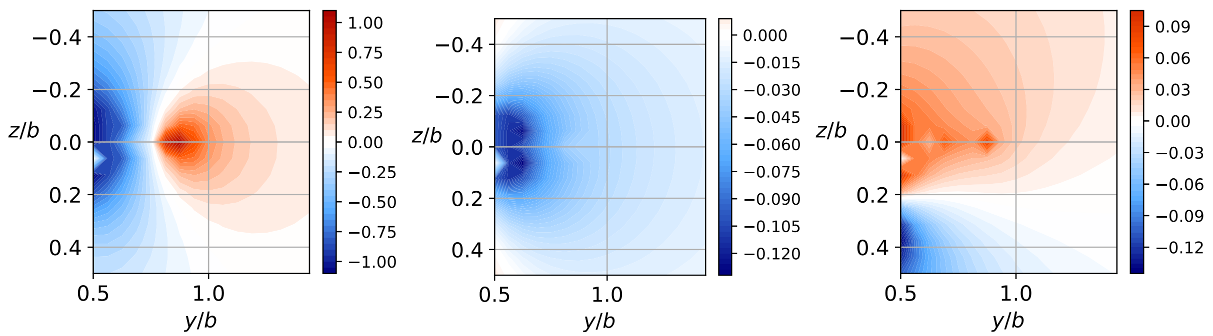
Acceleration differences

We display the acceleration differences $\Delta\dot{\mathbf{V}}_f = \dot{\mathbf{V}}_f - \dot{\mathbf{V}}_f^{isol}$ and $\Delta\dot{\boldsymbol{\omega}}_f = \dot{\boldsymbol{\omega}}_f - \dot{\boldsymbol{\omega}}_f^{isol}$ for different positions of a follower in the wake of a leader (located in $(0,0)$ in the (oy_iz_i) plane) in Fig. 4. The acceleration differences $\Delta\dot{U}_f$, $\Delta\dot{W}_f$ and $\Delta\dot{q}$ are symmetrical with respect to the vertical axis (oz_i), while the acceleration differences $\Delta\dot{V}_f$, $\Delta\dot{p}$ and $\Delta\dot{r}$ are assymmetrical with respect to that same axis.

The six graphs are obtained by trimming the control commands of the follower for each one of the positions in the (oy_iz_i) plane, while the leader is fixed, using the aerodynamics model described in Section II.A. The follower therefore maintains a steady configuration, as the aerodynamic loads induced by the control commands balance the effect of the incoming wake. This results in zero accelerations for the 6 DOF of the follower. However, the same aircraft using those same control commands, but without any external influence this time, experiences non-zero accelerations. This explains non-zeros acceleration differences in the graphs



(a) $\Delta\dot{U}_f \cdot \left(\frac{\rho U_\infty^2 S_w}{2m}\right)^{-1}$ (b) $\Delta\dot{V}_f \cdot \left(\frac{\rho U_\infty^2 S_w}{2m}\right)^{-1}$ (c) $\Delta\dot{W}_f \cdot \left(\frac{\rho U_\infty^2 S_w}{2m}\right)^{-1}$



(d) $\Delta\dot{p} \cdot \left(\frac{\rho U_\infty^2 S_w}{2mb}\right)^{-1}$ (e) $\Delta\dot{q} \cdot \left(\frac{\rho U_\infty^2 S_w}{2mb}\right)^{-1}$ (f) $\Delta\dot{r} \cdot \left(\frac{\rho U_\infty^2 S_w}{2mb}\right)^{-1}$

Figure 4 Contour plots of the differences in acceleration of the follower in the wake of the leader.

in Fig. 4. As an example, let us study the difference $\Delta\dot{U}_f$ for a position close to the optimal one in $(1, 0)$ in Fig. 4a; it is positive. When trimmed at that position, the follower benefits from an apparent drag reduction because positioned in the vicinity of the wake. The thrust required to maintain that position is therefore of lower amplitude than the one required to maintain a position in an unperturbed atmosphere. An aircraft using that thrust would therefore be subjected to a negative acceleration along that axis, which explains the positive difference. The asymmetry of $\Delta\dot{V}_f$ and $\Delta\dot{W}_f$ with respect to the horizontal axis $z = 0$ is explained by the asymmetry of the modeled aircraft, whose VTP is positioned above the HTP.

Convergence of the learning process

The results of the previous section indicate that relevant information can be obtained from acceleration differences and aircraft dynamics, in order to track the optimal position in a wake, without additional hardware instrumentation aboard compared to other tracking methods [11]. Indeed, a combination of those differences enables to map the environment around a wake. Associated to the actual speed of the follower, the control strategy will guide the latter towards an optimal position.

As mentioned in Section II.B, the difference in vertical acceleration $\Delta\dot{W}_f$ is chosen as the reward r_n as it has been proven to be close to the optimal flying position in the wake of a leader. Therefore, it cannot be used as an observation. We will not use the longitudinal acceleration $\Delta\dot{U}_f$ neither, as it is directly related to the drag which is difficult to measure directly in practice. On the contrary, the roll moment induced by a surrounding wake is known to be significant. The acceleration difference in roll $\Delta\dot{p}$ needs therefore to be used as an observation of the environment. We investigate the use of other measurements in addition to $\Delta\dot{p}$ in Fig. 5. This figure presents the statistical properties of the total return over the learning process: we evaluate the performance of twenty different instances of each combination.

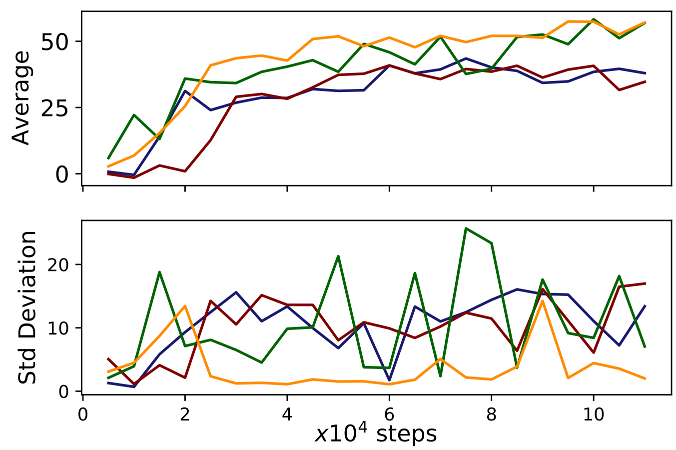


Figure 5 Training curves of the RL algorithm for different combinaisons of observations. (— : $[\Delta\dot{p}, \Delta\dot{r}]$, —: $[\Delta\dot{p}, \Delta\dot{V}_f]$, —: $[\Delta\dot{p}, \Delta\dot{r}, \Delta\dot{V}_f]$, —: $[\Delta\dot{p}, \Delta\dot{q}, \Delta\dot{r}, \Delta\dot{V}_f]$). The two graphs present the average and standard deviation of the total return g_n according to the number of learning steps performed.

We first investigate the use of one additional measurement, chosen between the lateral acceleration and yaw acceleration $\Delta\dot{q}$. Indeed unlike the pitch acceleration $\Delta\dot{q}$, they both show an asymmetry with the horizontal axis (oy_i), which enable to distinguish between an upper or a lower position compared to the leader (see Fig. 4). None of them gives better results compared to the other in terms of mean or standard deviation. The use of both of them in addition to $\Delta\dot{p}$ improves the average mean provided by the algorithm, while still keeping an important standard deviation. We finally choose to use a combination of all the remaining accelerations (i.e. $[\Delta\dot{p}, \Delta\dot{q}, \Delta\dot{r}, \Delta\dot{V}_f]$), that gives a high average return, while reducing the standard deviation.

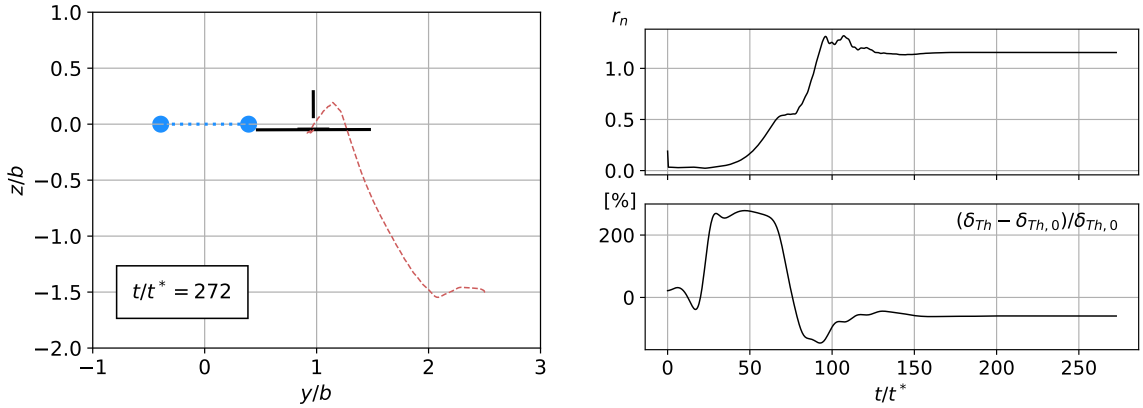
B. Tracking strategy behind a static leader

We apply the RL-based tracking strategy to a follower behind a fixed leader. We assume that the accelerations and dynamics of the follower can be effectively measured.

Fig. 6a is a visual representation of the convergence process performed by the follower towards the optimal position while Fig. 6b, Fig. 6c and Fig. 6d are time histories from that same simulation, with initial condition $(\frac{y_{f,0}}{b}, \frac{z_{f,0}}{b}) = (2.5, -1.5)$. The follower takes a period $t \simeq 150t^*$ (with $t^* = \frac{b}{U_\infty}$ the convective time), to reach a steady position, as clearly observed in Figs. 6c and 6d. We also observe the discontinuous reference speeds V_f^{ref} and W_f^{ref} that are computed by the trained neural net; they correspond to the action $\boldsymbol{a}_n$ taken by the aircraft. The update is performed at a rate $\Delta t_{\text{update}} = 6.8t^*$, which corresponds to the response time of the autopilot to a speed setpoint. The actual speeds V_f and W_f are therefore running with a lag close to this response time.

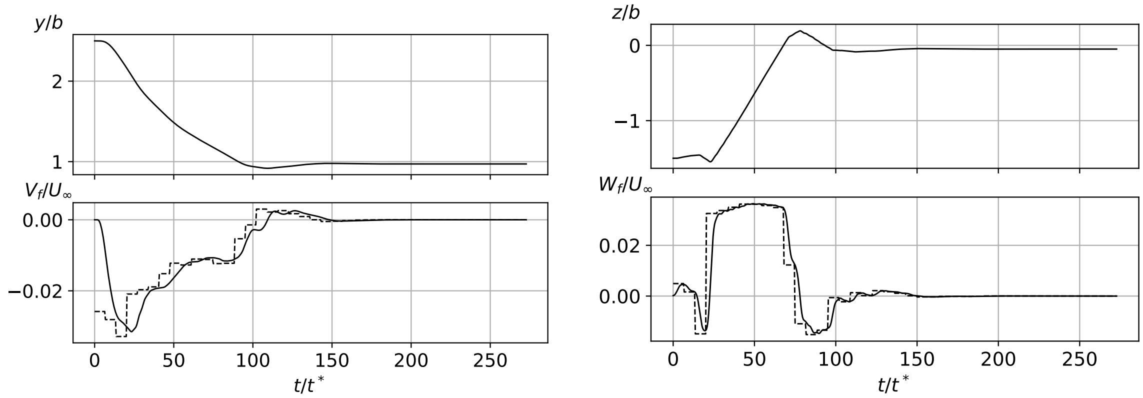
Fig. 6b presents the reward r_n received by the follower during its maneuver. We observe that the steady reward is not the maximum one the follower receives during the maneuver. This is explained by the second term of the definition of the reward in Eq. (7). This derivative term gives a higher reward when the follower takes an action leading to an increase of vertical acceleration $\Delta\dot{W}_f$ over one learning step Δt_{update}, which stimulates the convergence for distant positions compared to the leader. It also stimulate a rapid convergence towards the optimal position, and disappear when the follower reaches a steady position. Because the learning algorithm seeks to maximize the cumulative reward over time (and not the instantaneous one), we explain that the steady reward can be smaller than some observed during the transitional phase.

The best indication of the benefits of formation flight is also shown in Fig. 6b where we present the trust command δ_{Th} applied by the aircraft over time. It is compared to the thrust command $\delta_{Th,0}$ that the same



(a) Trajectory (- -) of a follower ( ) reaching an optimal position in the wake of a leader ( )

(b) Reward r_n received over time, and thrust command δ_{Th}



(c) Lateral action and resulting trajectory. The dashed line (- -) is the reference action V_f^{ref} while the solid line (—) is the actual speed V_f

(d) Vertical action and resulting trajectory. The dashed line (- -) is the reference action W_f^{ref} while the solid line (—) is the actual speed W_f

Figure 6 Performances and trajectory of a follower that converges towards the optimal position in the wake of a leader.

aircraft would have to applied in order to fly in a steady environment at the cruise speed U_∞, outside any wake. At the steady state, the model predicts a 60% thrust reduction at the optimal position. This reduction is overestimated compared to experimental results [1, 3] which have shown up to 20% apparent drag reduction. This overestimated prediction is explained by the aerodynamic model used that does not take into account any parasitic drag on the aircraft, but only the induced drag. The predicted thrust reduction thus only reflects the reduction in induced drag, and also does not take into account additional drag that could appear because of the wakes interactions. The even higher reduction observed at $t \simeq 90t^*$ is explained by the downward movement of the follower that benefits from the gravity to maintain the desired speed $\mathbf{V}_f$. Let us mention that a realistic aircraft would not enforce such a stiff control of its nominal speed U_∞, and would therefore not show such important excursions of the thrust command δ_{Th} as the ones observed on Fig. 6b.

C. Tracking strategy behind an oscillating leader

We now apply the exact same RL-based tracking strategy, this time for formation of four aircraft, where three followers track their predecessor wake and should therefore reproduce the oscillating maneuver performed by the leader. Fig. 7 presents snapshots of that formation over time.

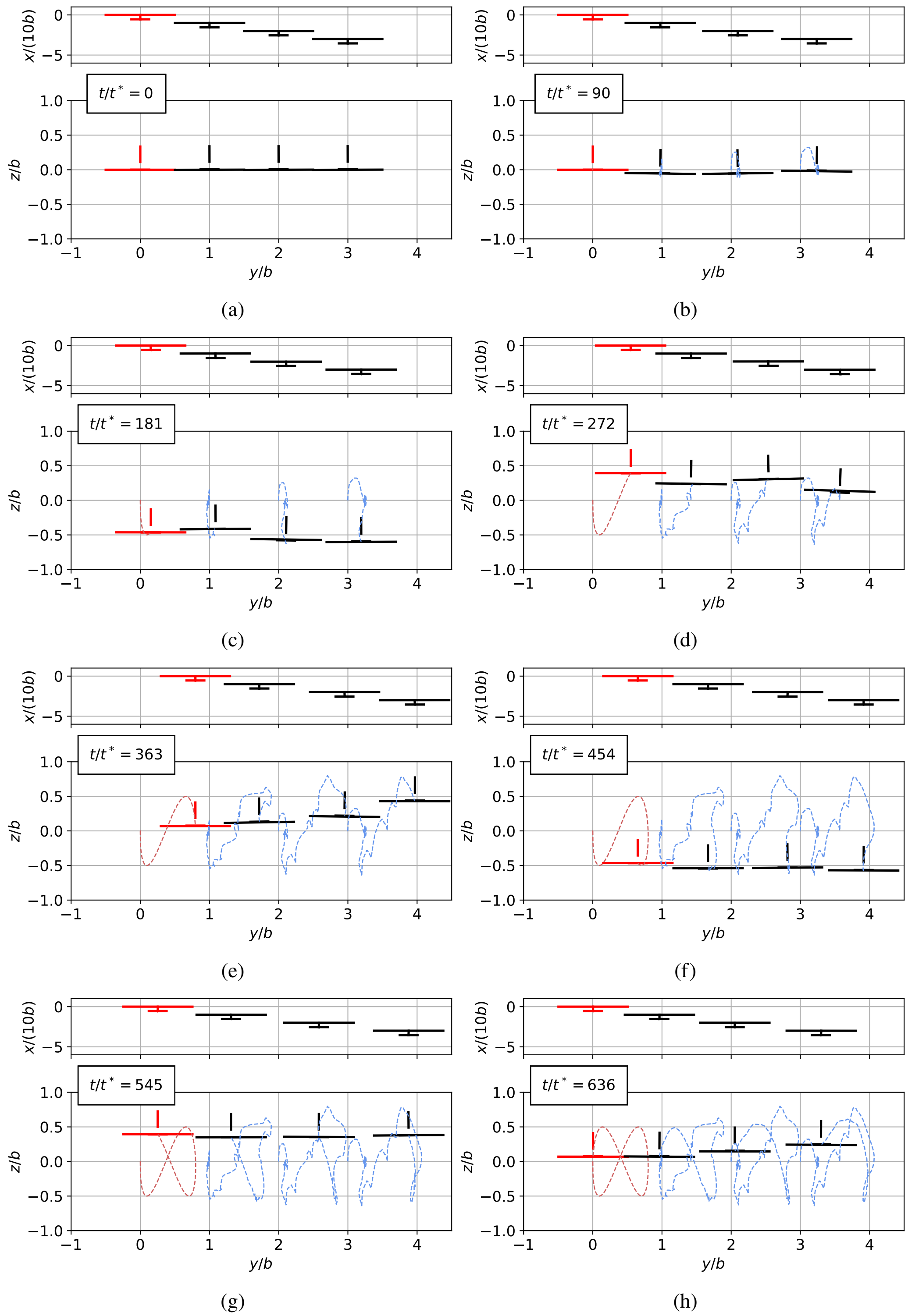
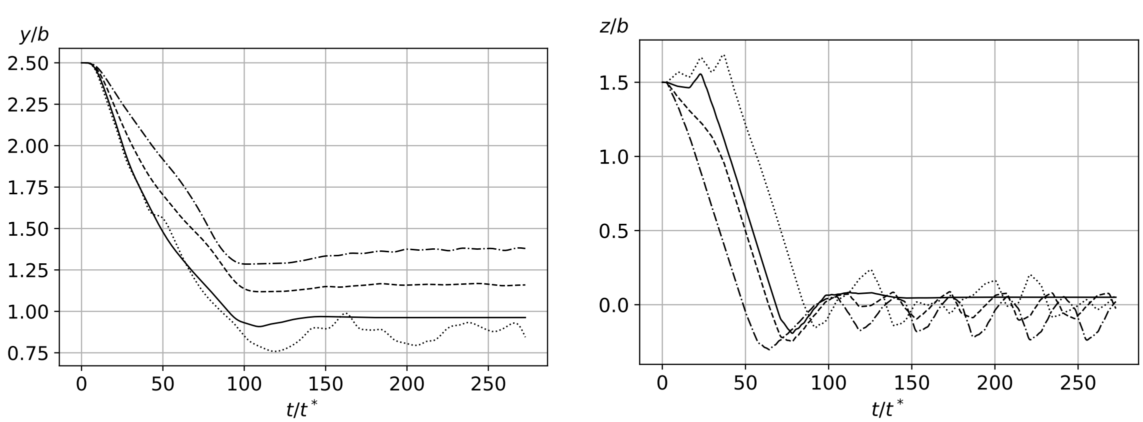


Figure 7 Evolution of a four-aircraft formation over time. The leader ( ) performed an oscillating maneuver while the followers ( ) use the RL-based controller in order to track the optimal position in the wake of their predecessor in the formation

We first observe that all the followers manage to keep the appropriate position in the wake of their predecessor. The RL-based controller is thus robust enough to deal with a complex environment for which it was not designed. Indeed, as described at the end of Section II.B, the learning process was performed assuming that the follower was flying in the wake of a single leader. While the first follower benefits from this favorable condition, the next two are influenced by more than one wake. This explains they do not reproduce the trajectory of the leader with the same accuracy as the first follower does. Also, the deviations performed by a follower tend to be reproduced or even amplified by the next one in the formation. Indeed, the wake produced by a predecessor is the result of its searching trajectory, which might lead to oscillations or "overshoots" as observed in Fig. 6a. The successor in the formation also behaves accordingly, and therefore tracks those tracking deviations. This is another explanation for the progressive degradation of the trajectory along the formation.

D. Results discussion

The learning strategy implemented in the previous sections assumes that the leader wake intensity $\Gamma_{0,L}$ is known to the follower, and remains constant over time (as described in Section II.B). Indeed, even if not directly provided to the follower, the acceleration maps presented in Fig. 4 are in a certain way "assimilated" by the follower during the exploration phase of the learning process. If $\Gamma_{0,L}$ changes (i.e. different leading aircraft, or progressive decrease over time), the intensity of the acceleration differences changes, which modifies those maps. The follower will act as if those maps had not changed, and the position he will reach will not be the optimal one for this new situation, as observed in Fig. 8. This simulation reproduces the exact same flight configuration as the one described in Section III.B, except for the the wake intensity $\Gamma_{0,L}$. It therefore shows the influence of this parameter on the trajectory of the follower, particularly on the steady position reached. In Fig. 8a, we observe that the follower tends to move away from the optimal position when the wake intensity increases, while it tends to line up with the leader wake when it decreases. This latter behavior is particularly dangerous, and unfavorable as the follower might enter the downwash area of the wake. The modification of $\Gamma_{0,L}$ also leads to an oscillatory behavior of the follower, as observed in Fig. 8a. Indeed, the tracking strategy has learned to converge towards specific values of the acceleration differences that map the environment. Now that this mapping has been modified, the strategy indefinitely searches for those values without ever reaching them.



(a) Lateral displacement

(b) Vertical displacement

Figure 8 Trajectory of followers impacted by wakes of various intensity (compared to the reference case $\Gamma_{0,L}^{ref}$) (—: $\Gamma_{0,L}/\Gamma_{0,L}^{ref} = 1$), (- - : $\Gamma_{0,L}/\Gamma_{0,L}^{ref} = 1.5$), (_._: $\Gamma_{0,L}/\Gamma_{0,L}^{ref} = 2$), (····: $\Gamma_{0,L}/\Gamma_{0,L}^{ref} = 0.7$)

While we have just highlighted the influence of the uncertainty of $\Gamma_{0,L}$ on the performances of the tracking strategy, similar analysis can be performed to all the parameters of the model used. In particular, even if not discussed in this work, the uncertainty on the "isolated aircraft model" (as described in Section II.B) would have a significative influence on the accuracy of the tracking strategy and would thus require further investigations.

IV. Conclusions

This study presents a bio-inspired wake tracking strategy that guides a follower to the optimal position in the wake of a leader in order to optimise the benefits of formation flight. To this end, a trained neural network provides velocity set-points to the autopilot of the follower that controls its 6 degrees-of-freedom dynamics. The resulting aerodynamic loads are computed thanks to a simplified, yet accurate aerodynamic model based on Prandtl's lifting line theory. The reinforcement learning framework is used to train the neural network based only on dynamic measurements of the follower. This ensures a total independence of the tracking method from direct measurements of the leader wake position.

After a convergence analysis of the RL algorithm with respect to the measurements used for the learning, two different flight configurations were studied. In the first one, we investigated the performance of the tracking strategy performed by a follower that flies towards the optimal position in the wake of a static leader. Rapid convergence and important drag reduction were observed. In the second one, we analysed the behavior of the tracking strategy when applied to four aircraft flying one behind the other. We showed that the tracking strategy was robust enough to deal with this more complex environment, even if inaccuracies were highlighted in terms of relative positioning of the aircraft. Indeed, we showed that the tracking strategy remained limited by its strong dependence on the dynamic properties of the follower and the properties of the leader wake (e.g. wake intensity). The parameters used during the learning process also oriented the performance of the tracking strategy, which is not desired for a robust application.

Further work will thus consist in providing the neural network with alternative observations of the environment, in order to no longer be dependent on a specific flight configuration, but rather be robust to complex environments. Dealing with uncertainties is another way of improving this first study of reinforcement learned wake tracking strategy. Thanks to the flexibility of the RL framework, and those promising results, such improvement can clearly be envisioned in the context of formation flight.

Acknowledgments

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement no. 725627).

The development work benefited from the computational resources provided by the supercomputing facilities of the Université catholique de Louvain (CISM/UCL) and the Consortium des Équipements de Calcul Intensif (CÉCI) en Fédération Wallonie Bruxelles (FWB) funded by the Fond de la Recherche Scientifique de Belgique (F.R.S.-FNRS) under Convention No. 2.5020.11.

References

[1] Beukenberg, M., and Hummel, D., "Aerodynamics, performance and control of airplanes in formation flight," *ICAS, Congress, 17 th, Stockholm, Sweden, Proceedings.*, 1990, pp. 1777–1794.

[2] Hummel, D., "The use of aircraft wakes to achieve power reductions in formation flight," *The Characterization & Modification of Wakes from Lifting Vehicles in Fluid*, Vol. AGARD-CP-584, NATO, 1996, pp. 1–13.

[3] Vachon, M., Ray, R., Walsh, K., and Ennix, K., "F/A-18 aircraft performance benefits measured during the autonomous formation flight project," Tech. Rep. TM-2003-210734, NASA, 2002. doi:10.2514/6.2002-4491.

[4] Bieniawski, S. R., Rosenzweig, S., and Blake, W. B., "Summary of Flight Testing and Results for the Formation Flight for Aerodynamic Benefit Program," *52nd Aerospace Sciences Meeting*, AIAA, 2014, pp. 1–26. doi:10.2514/6.2014-1457.

[5] Flanzer, T. C., Bieniawski, S. R., and Brown, J. A., "Advances in Cooperative Trajectories for Commercial Applications," *AIAA Scitech 2020 Forum*, 2020. doi:10.2514/6.2020-1257, URL `https://arc.aiaa.org/doi/abs/10.2514/6.2020-1257`.

[6] Ning, A., Flanzer, T. C., and Kroo, I. M., "Aerodynamic performance of extended formation flight," *Journal of Aircraft*, Vol. 48, No. 3, 2011, pp. 855–865. doi:10.2514/1.C031046.

[7] Blake, W., and Multhopp, D., "Design, performance and modeling considerations for close formation flight," *CLJ*, Vol. 150, 1998, p. 2.

[8] Hanson, C. E., Pahle, J., Reynolds, J. R., and Andrade, S., "Experimental Measurements of Fuel Savings During Aircraft Wake Surfing," *2018 Atmospheric Flight Mechanics Conference*, 2018, p. 3560. doi:10.2514/6.2018-3560.

[9] Michel, D. T., Dolfi-Bouteyre, A., Goular, D., Augère, B., Planchat, C., Fleury, D., Lombard, L., Valla, M., and Besson, C., "Onboard wake vortex localization with a coherent 1.5 µm Doppler LIDAR for aircraft in formation flight configuration," *Optics Express*, Vol. 28, No. 10, 2020, pp. 14374–14385. doi:10.1364/OE.377049, URL `http://www.opticsexpress.org/abstract.cfm?URI=oe-28-10-14374`.

[10] Rodenhiser, R. J., Durgin, W. W., and Johari, H., "Ultrasonic method for aircraft wake vortex detection," *Journal of aircraft*, Vol. 44, No. 3, 2007, pp. 726–732.

[11] Hemati, M. S., Eldredge, J. D., and Speyer, J. L., "Wake Sensing for Aircraft Formation Flight," *Journal of Guidance, Control, and Dynamics*, Vol. 37, No. 2, 2014, pp. 513–524. doi:10.2514/1.61114.

[12] Caprace, D.-G., Winckelmans, G., Chatelain, P., and Eldredge, J., "Wake Vortex Detection and Tracking for Aircraft Formation Flight," *AIAA Aviation 2019 Forum*, edited by AIAA, 2019. doi:10.2514/6.2019-3329, URL `https://arc.aiaa.org/doi/abs/10.2514/6.2019-3329`.

[13] Ransquin, I., Caprace, D.-G., and Chatelain, P., "Wake Vortex Detection and Tracking for Aircraft Formation Wake Vortex Detection and Tracking for Aircraft Formation Flight," , 2020. In preparation.

[14] Colvert, B., Alsalman, M., and Kanso, E., "Classifying vortex wakes using neural networks," *arXiv preprint*, 2017.

[15] Alsalman, M., Colvert, B., and Kanso, E., "Training bioinspired sensors to classify flows," *Bioinspiration & biomimetics*, Vol. 14, No. 1, 2018, p. 016009.

[16] Vlachas, P. R., Byeon, W., Wan, Z. Y., Sapsis, T. P., and Koumoutsakos, P., "Data-driven forecasting of high-dimensional chaotic systems with long short-term memory networks," *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, Vol. 474, No. 2213, 2018, p. 20170844.

[17] Coquelet, M., Bricteux, L., Lejeune, M., and Chatelain, P., "Biomimetic individual pitch control for load alleviation," *Journal of Physics: Conference Series*, Vol. 1618, No. 2, 2020, p. 022052.

[18] Verma, S., Novati, G., and Koumoutsakos, P., "Efficient collective swimming by harnessing vortices through deep reinforcement learning," *Proceedings of the National Academy of Sciences*, Vol. 115, No. 23, 2018, pp. 5849–5854. doi:10.1073/pnas.1800923115, URL `http://www.pnas.org/content/115/23/5849`.

[19] Reddy, G., Celani, A., Sejnowski, T. J., and Vergassola, M., "Learning to soar in turbulent environments," *Proceedings of the National Academy of Sciences*, Vol. 113, No. 33, 2016, pp. E4877–E4884.

[20] Cook, M. V., *Flight dynamics principles: a linear systems approach to aircraft stability and control*, Butterworth-Heinemann, 2012.

[21] Binetti, P., Ariyur, K. B., Krstic, M., and Bernelli, F., "Formation Flight Optimization Using Extremum Seeking Feedback," *Journal of Guidance, Control, and Dynamics*, Vol. 26, No. 1, 2003, pp. 132–142. doi:10.2514/2.5024.

[22] Chichka, D. F., Speyer, J. L., Fanti, C., and Park, C. G., "Peak-Seeking Control for Drag Reduction in Formation Flight," *Journal of Guidance, Control, and Dynamics*, Vol. 29, No. 5, 2006, pp. 1221–1230. doi:10.2514/1.15424, URL `https://doi.org/10.2514/1.15424`.

[23] Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S., "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," *Haarnoja, Tuomas and Zhou, Aurick and Abbeel, Pieter and Levine, Sergey*, 2018.

[24] Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al., "Tensorflow: A system for large-scale machine learning," *12th USENIX symposium on operating systems design and implementation (OSDI) 16)*, 2016, pp. 265–283.

[25] Hill, A., Raffin, A., Ernestus, M., Gleave, A., Kanervisto, A., Traore, R., Dhariwal, P., Hesse, C., Klimov, O., Nichol, A., Plappert, M., Radford, A., Schulman, J., Sidor, S., and Wu, Y., "Stable Baselines," https://github.com/hill-a/stable-baselines, 2018.