Learning to soar in turbulent environments

Gautam Reddy[a], Antonio Celani[b], Terrence J. Sejnowski[c,d,1], and Massimo Vergassola[a]

[a]Department of Physics, University of California, San Diego, La Jolla, CA 92093; [b]The Abdus Salam International Center for Theoretical Physics, I-34014 Trieste, Italy; [c]Howard Hughes Medical Institute, The Salk Institute for Biological Studies, La Jolla, CA 92037; and [d]Division of Biological Sciences, University of California, San Diego, La Jolla, CA 92093

Contributed by Terrence J. Sejnowski, April 28, 2016 (sent for review March 21, 2016; reviewed by Andrew G. Barto and Gregory Falkovich)

Birds and gliders exploit warm, rising atmospheric currents (thermals) to reach heights comparable to low-lying clouds with a reduced expenditure of energy. This strategy of flight (thermal soaring) is frequently used by migratory birds. Soaring provides a remarkable instance of complex decision making in biology and requires a long-term strategy to effectively use the ascending thermals. Furthermore, the problem is technologically relevant to extend the flying range of autonomous gliders. Thermal soaring is commonly observed in the atmospheric convective boundary layer on warm, sunny days. The formation of thermals unavoidably generates strong turbulent fluctuations, which constitute an essential element of soaring. Here, we approach soaring flight as a problem of learning to navigate complex, highly fluctuating turbulent environments. We simulate the atmospheric boundary layer by numerical models of turbulent convective flow and combine them with model-free, experience-based, reinforcement learning algorithms to train the gliders. For the learned policies in the regimes of moderate and strong turbulence levels, the glider adopts an increasingly conservative policy as turbulence levels increase, quantifying the degree of risk affordable in turbulent environments. Reinforcement learning uncovers those sensorimotor cues that permit effective control over soaring in turbulent environments.

thermal soaring | turbulence | navigation | reinforcement learning

Migrating birds and gliders use upward wind currents in the atmosphere to gain height while minimizing the energy cost of propulsion by the flapping of the wings or engines (1, 2). This mode of flight, called soaring, has been observed in a variety of birds. For instance, birds of prey use soaring to maintain an elevated vantage point in their search for food (3); migrating storks exploit soaring to cover large distances in their quest for greener pastures (4). Different forms of soaring have been observed. Of particular interest here is thermal soaring, where a bird gains height by using warm air currents (thermals) formed in the atmospheric boundary layer. For both birds and gliders, a crucial part of the thermal soaring is to identify a thermal and to find and maintain its core, where the lift is typically the largest. Once migratory birds have climbed up to the top of a thermal, they glide down to the next thermal and repeat the process, a migration strategy that strongly reduces energy costs (4). Soaring strategies are also important for technological applications, namely, the development of autonomous gliders that can fly large distances with minimal energy consumption (5).

Thermals arise as ascending convective plumes driven by the temperature gradient created due to the heating of the earth's surface by the sun (6). Hydrodynamic instabilities and processes that lead to the formation of a thermal inevitably give rise to a turbulent environment characterized by strong, erratic fluctuations (7, 8). Birds or gliders attempting to find and maintain a thermal face the challenge of identifying the potentially long-lived and large-scale wind fluctuations amid a noisy turbulent background. The structure of turbulence is highly complex, with fluctuations occurring at many different scales and long-ranged correlations in space and time (9, 10). We thereby expect nontrivial correlations between the large-scale convective plumes and the locally fluctuating quantities. Thermal soaring is a particularly interesting example of navigation within turbulent flows, because the velocity amplitudes of a glider or bird are of the same order of magnitude as the fluctuating flow they are immersed in.

It has been frequently observed and attested by glider pilots that birds are able to identify and navigate thermals more accurately than human pilots endowed with modern instrumentation (11). It is an open problem, however, what sensorimotor cues are available to birds and how they are exploited, which constitutes a major motivation for the present study.

An active agent navigating a turbulent environment has to gather information about the fluctuating flow while simultaneously using the flow to ascend. Thus, the problem faced by the agent bears similarities to the general problem of balancing exploration and exploitation in uncertain environments, which has been well studied in the reinforcement learning framework (12). The general idea of reinforcement learning is to selectively reinforce actions that are highly rewarding and thereby have the reinforced actions chosen when the situation reoccurs. The solution to a reinforcement learning problem typically yields a behavioral policy that is approximately optimal, where optimality is defined in the sense of maximizing the reward function used to train the agent.

The previous description suggests that reinforcement learning methods are poised to deliver effective strategies of soaring flight. Past applications are indeed promising, yet they have considered the soaring problem in unrealistically simplified situations, with no turbulence or with fluctuations modeled as Gaussian white noise. Ref. 13 considered the learning problem associated with finding the center of a stationary thermal without turbulence, and used a neural-based algorithm to recover the empirical rules proposed by Reichmann (14) to locate the core of the thermal. Other attempts (15, 16) have used neural networks and Q-learning to find strategies to center a turbulence-free thermal. Akos et al. (17) show that these simple rules fail even in the presence of modest velocity fluctuations modeled as Gaussian white noise, and express the need for strategies that could work in realistic turbulent flows.

Significance

Thermals are ascending currents that typically extend from the ground up to the base of the clouds. Birds and gliders piggyback thermals to fly with a reduced expenditure of energy, for example, during migration, and to extend their flying range. Flow in the thermals is highly turbulent, which poses the challenge of the orientation in strongly fluctuating environments. We combine numerical simulations of atmospheric flow with reinforcement learning methods to identify strategies of navigation that can cope with and even exploit turbulent fluctuations. Specifically, we show how the strategies evolve as the level of turbulent fluctuations increase, and we identify those sensorimotor cues that are effective at directing turbulent navigation.

Author contributions: G.R., A.C., T.J.S., and M.V. designed research; G.R. performed research; G.R., A.C., and M.V. analyzed data; and G.R., A.C., T.J.S., and M.V. wrote the paper.

Reviewers: A.G.B., University of Massachusetts; and G.F., Weizmann Institute of Science.

The authors declare no conflict of interest.

[1]To whom correspondence should be addressed. Email: terry@salk.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1606075113/-/DCSupplemental.

Here, we enforce realistic aerodynamic constraints on the flight of gliders and train them in complex turbulent environments by using reinforcement learning algorithms. We show that the glider finds an effective strategy for soaring, and we identify sensorimotor cues that are most relevant for guiding turbulent navigation. Our soaring strategy is effective even in the presence of strong fluctuations. The predicted strategy of flight lends itself to field experiments with remote-controlled gliders and to comparisons with the behavior of soaring birds.

Models

We first describe the models used for the simulation of the atmospheric boundary layer flow, the mechanics of flight, and the reinforcement learning algorithms that we have used. The next section will then present the corresponding results.

Modeling the Turbulent Environment. Conditions ideal for thermal soaring typically occur during a sunny day, when a strong temperature gradient between the surface of the Earth and the top of the atmospheric boundary layer creates convective thermals (7, 8). The soaring of birds and gliders primarily occurs within this convective boundary layer. The mechanical and thermal forces within the boundary layer generate turbulence characterized by strongly fluctuating wind velocities.

Key physical aspects of the flow in the convective boundary layer are governed by Rayleigh–Bénard convection (see ref. 9 for a review). The corresponding equations are derived from the Navier–Stokes equations with coupled temperature and velocity fields simplified using the Boussinesq approximation. The dimensionless Rayleigh–Bénard equations read as follows:

$$\frac{\partial \boldsymbol{u}}{\partial t}+\boldsymbol{u}\cdot\nabla\boldsymbol{u}=-\nabla P+\left(\frac{\mathrm{Pr}}{\mathrm{Ra}}\right)^{1/2}\nabla^2\boldsymbol{u}+\theta\hat{z}, \qquad [1]$$

$$\frac{\partial \theta}{\partial t}+\boldsymbol{u}\cdot\nabla\theta=\frac{1}{(\mathrm{Pr}\ \mathrm{Ra})^{1/2}}\nabla^2\theta, \qquad [2]$$

where $\boldsymbol{u}, \theta$, and P are the velocity, temperature, and pressure fields, respectively. The vertical direction coincides with the z axis. The temperature appears in the dynamics of the velocity field as a buoyant forcing term. The equations contain two dimensionless quantities that determine the qualitative behavior of the flow: the Rayleigh number, Ra, and the Prandtl number, Pr. When Ra is beyond a critical value $\sim 10^3$, the thermally generated buoyancy drives the flow toward instability. In this regime, the flow is characterized by large-scale convective cells and turbulent eddies at every length scale. In the atmosphere, the Rayleigh number can reach up to Ra = 10^{15} to 10^{20}. In such high-Rayleigh number regimes, the flow is strongly turbulent and numerical simulations of convection in the atmosphere are thus plagued by the same limitations of simulating fully developed turbulent flows. We performed direct numerical simulations of Rayleigh–Bénard convection at Ra = 10^8 using the Gerris Flow Solver (18) (see *Supporting Information* for more details about the grid and the numerical scheme). Our test arena is a 3D cubical box of side length 1 km in physical units. We impose periodic boundary conditions on the lateral walls and no-slip on the floor and the ceiling of the box. The floor is fixed at a high temperature (which is rescaled to $\theta = 1$), and the ceiling is fixed at $\theta = 0$.

A small, random perturbation in the flow quickly leads to an instability and to the formation of coherent thermal plumes within the chamber. Snapshots of the velocity and temperature fields at the statistically stationary state are shown in Fig. 1*A*. The statistical properties of the flow are consistent with those observed in previous works (19, 20), particularly the Nusselt number (which measures the ratio of convective to conductive heat transfer) and the mean temperature and velocity field profiles (Fig. S1).

To test the robustness of our learned policies of flight with respect to the modeling of turbulence, we also considered an alternative to the Rayleigh–Bénard flow. Specifically, we considered a kinematic model of turbulence that extends the one in ref. 21 to the inhomogeneous case relevant for the atmospheric boundary layer (*Methods*). Results for the kinematic model confirm the robustness of our conclusions and the learned policy has similar features in both flows (*Supporting Information* and Figs. S2–S4). Below, we shall focus on the simulations of the Rayleigh–Bénard flow described above.

Glider Mechanics. A bird or glider flying in the flow described above with a fixed, stretched-out wing is safely assumed to be in mechanical equilibrium, except for centripetal forces while turning (22, 23). A glider with weight W traveling with velocity $\boldsymbol{v}$ experiences a lift force L perpendicular to its velocity and a drag force D antiparallel to its velocity (see Fig. 1*C* for a force body diagram). The glider has no engine and thus generates no thrust. The magnitudes of the lift and the drag depend on the speed v, the angle of attack α, the density of air ρ and the surface area S of the wing as follows: $L = (1/2)\rho S v^2 C_L(\alpha)$ and $D = (1/2)\rho S v^2 C_D(\alpha)$. The glide angle γ, which is the angle between the velocity and its projection on the horizontal, determines the ratio of the climb rate $v_c(<0)$ and the horizontal speed $v_\perp$. Balancing the forces on the glider, and accounting for the centripetal acceleration, the velocity of the glider and its turning rate are obtained as follows:

$$\tan\gamma=\frac{-v_c}{v_\perp}=\frac{D}{L\cos\mu}=\frac{C_D(\alpha)}{C_L(\alpha)\cos\mu}; \qquad [3]$$

$$\ddot{y}=g\cos\gamma\tan\mu;\quad v^2=\frac{2mg\sin\gamma}{\rho S C_D(\alpha)}. \qquad [4]$$

Here, $\ddot{y}$ is the centripetal acceleration. The ratio mg/S is called the wing loading of the glider (22). The kinematics of a glider is therefore set by the wing loading and the dependence of the lift and the drag coefficients on the angle of attack. The general features of the lift and drag coefficient curves for a typical symmetric airfoil are described in ref. 24; the resulting dependence of the velocity on the angle of attack is shown in Fig. 1*B*. The glider can be maneuvered by controlling the angle of attack, which changes the speed and climb rate of the glider, or by banking the glider to turn.

The Learning Algorithm. To identify effective strategies of soaring flight in turbulent flows, we used the reinforcement learning algorithm state–action–reward–state–action (SARSA) (12). Historically, the algorithm was inspired by the theory of animal learning, and its model-free nature allows for learning previously unknown strategies driven by feedback on performance (25).

Reinforcement learning problems are typically posed in the framework of a Markov decision process (MDP). In an MDP, the agent traverses a state space with transition probabilities that depend only on the current state s and the immediate next state s', as for a Markov process. The transition probabilities can be influenced by taking actions at each time step. After every action, the agent is given some reward $r(s, s', a)$, which depends on the states s and s' and the chosen action a. The ultimate goal of reinforcement learning algorithms is to find the optimal policy π^*, that is, to find the probability of choosing action a given the state s. The optimal policy maximizes for each state s the sum of discounted future rewards $V_{\pi_s^a}(s) = \langle r_0\rangle + \beta\langle r_1\rangle + \beta^2\langle r_2\rangle + \ldots$, where $\langle r_i\rangle$ is the expected reward after i steps, β is the discount factor ($0 \le \beta 1$), and the sum above obviously depends on the policy π_s^a. When β is close to zero, the optimal policy greedily maximizes the expected immediate reward, leading to a purely exploitative strategy. As β gets closer to unity, later rewards

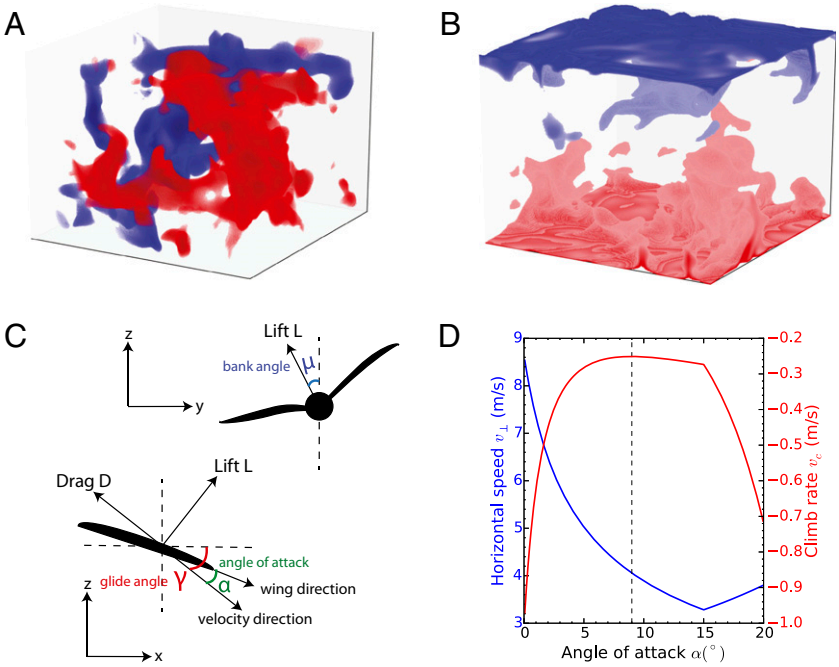

Fig. 1. Snapshots of the vertical velocity (*A*) and the temperature fields (*B*) in our numerical simulations of 3D Rayleigh–Bénard convection. For the vertical velocity field, the red and blue colors indicate regions of large upward and downward flow, respectively. For the temperature field, the red and blue colors indicate regions of high and low temperature, respectively. Notice that the hot and cold regions drive the upward and downward branches of the convective cell, in agreement with the basic physics of convection. (*C*) The force-body diagram of flight with no thrust, that is, without any engine or flapping of wings. The figure also shows the bank angle μ (blue), the angle of attack α (green), and the glide angle γ (red). (*D*) The range of horizontal speeds and climb rates accessible by controlling the angle of attack. At small angles of attack, the glider moves fast but also sinks fast, whereas at larger angles, the glider moves and sinks more slowly. If the angle of attack is too high, at about 16°, the glider stalls, leading to a sudden drop in lift. The vertical black dashed line shows the fixed angle of attack for most of the simulations (*Results, Control over the Angle of Attack*).

contribute significantly and more exploratory strategies are preferred.

The SARSA algorithm finds the optimal policy by estimating for every state–action pair its Q function defined as the expected sum of future rewards given the current state s and the action a. At each step, the Q function is updated as follows:

$$Q(s,a) \to Q(s,a) + \eta(r + \beta Q(s',a') - Q(s,a)), \qquad [5]$$

where r is the received reward and η is the learning rate. The update is made online and does not require any prior model of the flow or the flight. This feature is particularly relevant in modeling decision-making processes in animals. When the algorithm is close to convergence, the Q function approaches the solution to Bellman's dynamic programming equations (12). The policy π_s^a, which encodes the probability of choosing action a at state s, approaches the optimal one π^* and is obtained from the Q function via a Boltzmann-like expression:

$$\pi_s^a \propto \exp\left(-\hat{Q}(s,a)/\tau_{\text{temp}}\right), \qquad [6]$$

$$\hat{Q}(s,a) = \frac{\max_{a'} Q(s,a') - Q(s,a)}{\max_{a'} Q(s,a') - \min_{a'} Q(s,a')}. \qquad [7]$$

Here, τ_{temp} is an effective "temperature": when $\tau_{\text{temp}} \gg 1$, actions are only weakly dependent on the associated Q function ; conversely, for τ_{temp} small, the policy greedily chooses the action with the largest Q. The temperature parameter is initially chosen large and lowered as training progresses to create an annealing effect, thereby preventing the policy from getting stuck in local extrema. Parameters used in our simulations can be found in Table S1.

In the sequel, we shall qualify the policy identified by SARSA as optimal. It should be understood, however, that the SARSA algorithm (as other reinforcement learning algorithms) typically identifies an approximately optimal policy and "approximately" is skipped only for the sake of conciseness.

Results

Sensorimotor Cues and Reward Function for Effective Learning. Key aspects of the learning for the soaring problem are the sensorimotor cues that the glider can sense (state space) and the choice of the reward used to train the glider to ascend quickly. As the state and action spaces are continuous and high-dimensional, it is necessary to discretize them, which we realize here by a standard lookup table representation. The height ascended per trial, averaged over different realizations of the flow, serves as our performance criterion.

The glider is allowed control over its angle of attack and its bank angle (Fig. 1*B*). Control over the angle of attack features two regimes: (*i*) at small angles of attack, the horizontal speed is large and the climb rate is small (the glider sinks quickly); (*ii*) at large angles of attack but below the stall angle, the horizontal speed is small, whereas the climb rate is large. The bank angle controls the heading of the glider, and we allow for a range of variation between −15° and 15°. Exploring various possibilities, we found that three actions are minimally sufficient: increasing, decreasing, or preserving the angle of attack and the bank angle. The angle of attack and bank angle were incremented/decremented in steps of 2.5° and 5°, respectively. In summary, the glider can choose 3^2 possible actions to control its navigation in response to the sensorimotor cues described hereafter.

Our rationale in the choice of the state space was trying to minimize biological or electronic sensory devices necessary for control. We tested different combinations of local sensorimotor

cues that could be indicative of the existence of a thermal. These were the vertical wind velocity u_z, the vertical wind acceleration a_z, torques τ, the local temperature θ, and their 16 possible combinations. Namely, if u denotes the local wind speed, we define the wind acceleration as $a_z = (u_z^{(t)} - u_z^{(t-1)})/\Delta t$ and the "torques" as $\tau = (u_{z+} - u_{z-})l$, where u_{z+} and u_{z-} are the vertical wind velocities at the left and the right wing, l is the wingspan of the glider, and Δt is the step used for time discretization (see below). After experimentation with various architectures, we found that a lookup table structure with three states per observable, corresponding to positive high, negative high, and small values, ensures good performance. The chosen thresholds, a_z^{thresh} and τ^{thresh}, that demarcate the large and small values in our scheme are listed in Table S1.

As for the reward function, we found that a purely global reward, that is, awarded at the end of a trial without any local guidance, does not propagate easily to early state–action pairs for realistically long trials. Eligibility traces (12), which maintain a memory of past state–action pairs and their rewards, did not alleviate the issue. For gliders or migrating birds, a fall can be extremely disadvantageous, and we account for this by having a glider that touches the surface receive a large negative reward as a penalty. After a broad exploration of various choices, we heuristically found that best soaring performances are obtained by a local-in-time reward that linearly combines the vertical wind velocity and the wind acceleration achieved at the subsequent time step, that is, $R = u_z + Ca_z$ (see Table S1 for the chosen value). We observe that performance does not change significantly for a wide range of values of C.

Flight Training. The glider is first trained on a set of trials and its performance is then tested on 500 trials. Trials consist of independent statistical realizations of the turbulent flow. The glider flight is discretized by time steps $\Delta t = 1$ s, which is an estimate for the control times of the glider and the timescales of the turbulent eddies at the size of the glider. Each trial lasts for 2.5 min, which is roughly one-half the relaxation time of the large-scale convective flow at steady state. The duration captures the order of magnitude of the typical time, ~ 10 min, for birds to reach the base of the clouds.

The velocity relative to the ground of the glider is $u + v$, where u and v are the contributions due to the wind and the glider velocity, respectively. If u_{rms} is the root-mean-squared speed of the flow and v_{glider} is the typical airspeed of the glider, we introduce their dimensionless ratio $\hat{u}_{\text{rms}} = u_{\text{rms}}/v_{\text{glider}}$. At small $\hat{u}_{\text{rms}}$, fluctuations are weak. Conversely, at large $\hat{u}_{\text{rms}}$, the glider has less time to react to rapidly changing velocities; that is, the environment is strongly fluctuating. Moreover, in that regime, the glider is carried away by the flow and the amount of control the glider has over its trajectory is reduced. We expect that the policy of flight learned by the glider will differ between the regimes of weak and strong fluctuations.

Learning in Different Flow Regimes. A qualitative sense of the efficiency of the training in a fluctuating regime is illustrated in Fig. 2. The trajectories go from random paths to the spirals that are characteristic of the thermal soaring flights of birds and gliders. Fig. 3*A* quantifies the significant improvement in performance due to training and shows that training for a few hundred trials suffices for convergence with negligible overfitting for larger training sets. To compare performance in flows of different mean speeds, we train and test gliders in flows with varying u_{rms}. Fig. 3*B* shows the gain in height as a function of $\hat{u}_{\text{rms}}$. As expected, we observe two regimes: (*i*) for weak and moderate fluctuations, $\hat{u}_{\text{rms}} \lesssim 1$, the gain in height follows a rapidly increasing trend; (*ii*) for strong fluctuations, $\hat{u}_{\text{rms}} \gtrsim 1$, gains still increase but more slowly. Because the ascended height depends on the flow speed, Fig. 3*B* also shows the soaring efficiency χ, defined as the difference between $\Delta h(\hat{u}_{\text{rms}})$ and $\Delta h(0)$ divided by $w_{\text{rms}}\Delta T$, where w_{rms} is the rms vertical speed of the flow and $\Delta T = 150$ s is the duration of a trial (*Supporting Information* and Fig. S1 for the value of w_{rms}). If the glider did not attempt to selectively find upward currents, χ would vanish, whereas $\chi = 1$ corresponds to a glider perfectly capturing vertical currents. As the flow speed increases, the efficiency shows a downward trend that reflects the increasing difficulty in control due to higher levels of fluctuations.

The performance of different gliders soaring simultaneously within the same flow does not vary significantly, indicating that an ensemble of gliders learn a uniquely optimal policy. The performance over different realizations for a single glider varies wildly, with a SD of the final height of the same magnitude as the final height itself when $\hat{u}_{\text{rms}} \approx 1$. Despite this wide variation, the number of failures (i.e., the glider touches the ground) always decreases rapidly to almost zero with the number of training trials.

Role of Wind Acceleration and Torques. Our learning procedure allows us to test the possible local sensorimotor cues that give good soaring performance. For each cue, we define a mean level and upper and lower thresholds symmetrically around the mean value. The performance was found to be largely independent of the chosen thresholds.

In Fig. 3*C*, we show a comparison between the performance of a few different combinations of the cues. We found that the pairing of vertical wind acceleration and torques, gauged in terms of the average height ascended per trial, works best (results in Fig. 3 *A* and *B* are obtained using this pair). Intuitively,

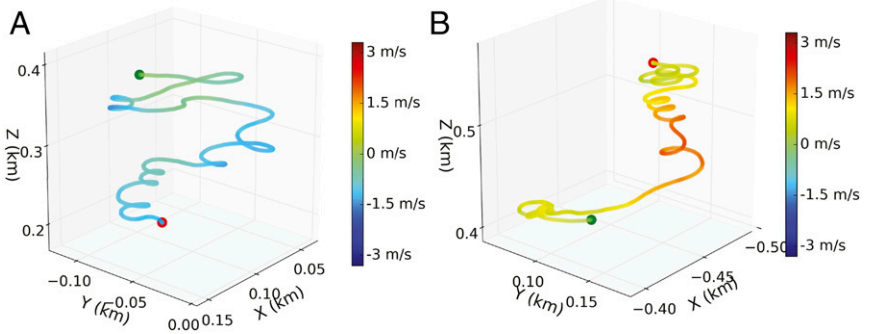

Fig. 2. Typical trajectories of an untrained (*A*) and a trained (*B*) glider flying within a Rayleigh–Bénard turbulent flow, as shown in Fig. 1. The colors indicate the vertical wind velocity experienced by the glider. The green and red dots indicate the start and the end points of the trajectory, respectively. The untrained glider takes random decisions and descends, whereas the trained glider flies forming the characteristic spiraling patterns in regions of strong ascending currents, as observed in the thermal soaring of birds and gliders (see, e.g., figure 2 in ref. 11).

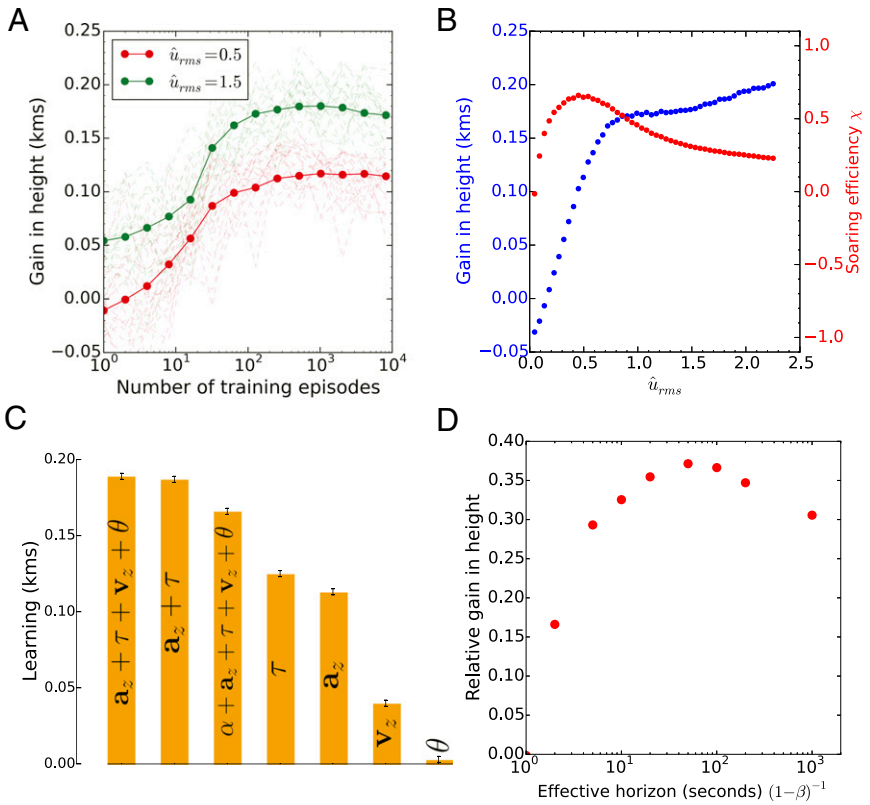

Fig. 3. The soaring performance of flight policies and the sensed sensorimotor cues. (*A*) The learning curve for two different turbulent fluctuation levels, as quantified by the ratio $\hat{u}_{\rm rms}$ of the rms variations of the flow and the airspeed of the glider. The two values $\hat{u}_{\rm rms}=0.5$ (red) and $\hat{u}_{\rm rms}=1.5$ (green) show the increase in the average ascended height per trial with the size of the training set. The training saturates after ≈ 250 trials. The green and red dotted lines show the learning curves of 20 individual gliders. (*B*) The panel shows the average height ascended for different $\hat{u}_{\rm rms}$ (blue). We also plot the soaring efficiency $\chi(\hat{u}_{\rm rms})$ as defined in the text. The efficiency takes into account the stronger ascending velocities that are a priori available when $\hat{u}_{\rm rms}$ increases. The difficulty is of course that higher velocities are also associated to stronger fluctuations. The efficiency indeed shows a downward trend that reflects the increasing difficulty in control as fluctuations increase. (*C*) A comparison of the average gain in height for different combinations of sensorimotor cues (vertical acceleration $\mathbf{a}_z$ and velocity $\mathbf{v}_z$, torque τ and temperature θ) sensed by the glider. Vertical wind velocities and temperature also give minor contributions compared with the performance of vertical wind acceleration and torque. The third bar includes the performance when the control of the angle of attack α is included as a possible action. The contribution is marginal and the convergence is actually slowed down so that the final performance after a finite number of training trials is slightly inferior to the first bar. The error bars show the 95% confidence interval of the average gain in height. (*D*) The relative improvement in height gained with respect to a greedy strategy, that is, with discount factor $\beta=0$. A reinforcement learning policy that is not greedy, that is, $\beta\neq 0$, has significantly better performance, demonstrating that long-term planning improves soaring. For *A*, *B*, and *D*, the error bars are smaller than the symbol size and are not shown.

the combination of vertical wind acceleration and torques provides information on the gradient of the vertical wind velocity in two complementary directions, thus allowing the glider to decide between turning or continuing along the same path. Conversely, the vertical wind velocity does indicate the strength of a thermal, but it does not guide the glider to the core of the thermal. The pair acceleration and torque allows the glider to climb the thermal toward the core and also detect the edge of a thermal so that the glider can stay within the core. The resulting pattern within a thermal is a spiral that occurs solely from actions based on local observables and minimal memory use. Temperature fails to improve performance, which could be intuited as the temperature field is highly intermittent and is itself a convoluted function of the turbulent velocity (26, 27).

Control over the Angle of Attack. Fig. 3*C* shows that control over the angle of attack does not influence significantly the performance in climbing an individual thermal. The angle of attack should play an important role, however, in other situations, namely, during cross-country races or bird migration, where gliders need to cover large horizontal distances and control over the horizontal speed and sink rate is needed (11, 28, 29). To verify this expectation, we considered a simple test case of a glider traversing, without turning, a 2D track consisting of a series of ascending or descending columns of air with turbulence added on top. We found that control over the angle of attack indeed improves the gain in height (*Supporting Information* and Fig. S5) and the glider learns to increase its pace during phases of descent while slowing down during periods of ascending currents. We expect that the differing roles of the angle of attack for soaring between and within thermals holds true for birds as well, a prediction that can be tested in field experiments.

In the sequel, we shall analyze the soaring in a single thermal. We fix then for simplicity the angle of attack at $\sim 9°$ (where the climb rate is the largest; Fig. 1*B*), and the pair acceleration–torque as sensorimotor cues sensed by the glider (Fig. 3*C*).

Dependence on the Temporal Discounting. The performance of the glider as a function of the temporal discount factor β is shown in Fig. 3*D*. The gain in height increases as the effective time horizon $(1-\beta)^{-1}$ grows, reaches a maximum at ≈ 100 s, and then slowly declines. The best time horizon is comparable with the timescale of the flow patterns at the height reached by the glider. This demonstrates that long-term planning is crucial for soaring

and the importance of a relatively long-term strategy to effectively use the ascending thermals.

Optimal Flight Policy. The Q function learned by the SARSA algorithm defines the optimal state–action policy via Eq. **6**. An optimal policy associates the choice of an action to the pair acceleration–torque (a_z, τ). The optimal action is chosen among the three options: (*i*) increase the bank angle μ by 5°; (*ii*) decrease μ by 5°; (*iii*) keep μ unchanged. In Fig. 4*A*, we show a comparison between the policy for the two regimes of weak and strong fluctuations.

The policies in Fig. 4 have a few intuitive features that are preserved at different flow speeds. For instance, when the glider experiences a negative wind acceleration, the optimal action is to sharply bank toward the side of the wing that experiences larger lift. When the glider experiences a large positive acceleration and no torque, the glider continues flying along its current path. Despite these similarities, the policies exhibit marked differences, which we proceed to analyze.

For each a_z, τ pair, it is useful to consider its preferred angles (the green circles in Fig. 4), that is, those angles that the policy leads to if the pair a_z, τ is maintained fixed. We observe that the preferred bank angles of gliders trained in a strong flow are relatively moderate, and the policy in general is more conservative. Consider, for instance, the case of zero torque and zero acceleration (column 5 of the policies in Fig. 4). The optimal bank action in the weak flow regime is to turn as much as possible, in contrast to the policy in the strong flow regime, which is to not turn. Another interesting qualitative difference is when the glider experiences negative acceleration and significant torque on the right wing (column 1 of the policies in Fig. 4). In the weak flow regime, if the glider is already banked to the left (negative bank angles), the policy is to bank further left to complete a full circle. In the strong flow regime, the policy is once again more conservative, preferring to not risk the full turn.

A policy becoming more conservative and risk averse as fluctuations increase is consistent with the balance of exploration and exploitation (12). In a noisy environment, where a wrong decision can lead to highly negative consequences, we expect an active agent to play safe and tend to gather more information before taking action. In a turbulent environment, we expect the glider to exploit (avoid) only significantly large positive (negative) fluctuations along its trajectory while filtering out transient, small-scale fluctuations. In the next subsection, we shall further confirm this expectation by tracking the changes in the optimal policy with the flow speed and extracting a few general principles of the optimal flight policy.

Optimal Bank Angles. To quantify the description of the optimal policy shown in Fig. 4*A*, we consider the distributions of the bank angle μ given the acceleration a_z and torque τ in the previous time step, that is, $\Pr(\mu^{(t+1)}|a_z^{(t)}, \tau^{(t)})$. We define the optimal bank angle as follows:

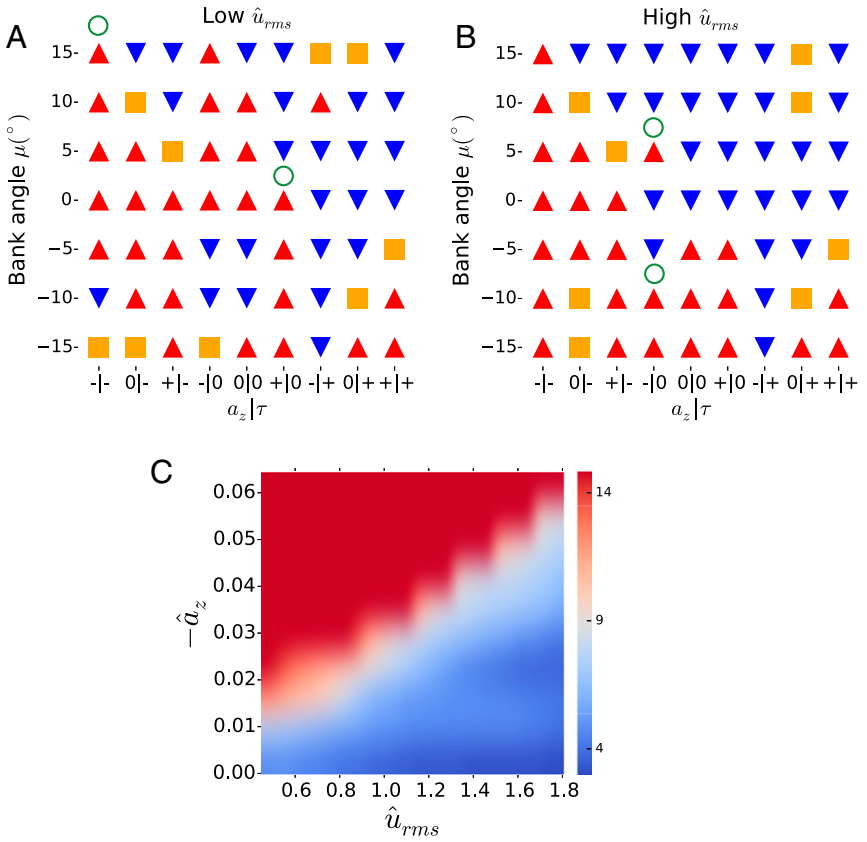

Fig. 4. Policies of flight for different levels of turbulent fluctuations. (*A*) $\hat{u}_{\text{rms}} = 0.5$ and (*B*) $\hat{u}_{\text{rms}} = 1.5$, with $\hat{u}_{\text{rms}}$ defined as in Fig. 3. The plot shows the optimal action on the bank angle μ upon receiving a given sensorimotor cue of vertical acceleration and torque (a_z, τ). Here +, −, and 0 denote positive high, negative high, and low values for a_z and τ as discussed in the text. The red upward arrow, blue downward arrow, and orange square indicate that the optimal policy is to increase, decrease, or maintain the same bank angle, respectively. A few instances of the preferred angles that one eventually reaches by maintaining (a_z, τ) fixed are denoted by green circles. Note that the policy at $\hat{u}_{\text{rms}} = 1.5$ is more conservative compared with that at 0.5; namely, preferred angles are smaller for the former. (*C*) A heat map showing the optimal bank angle (Eq. **8**) at a particular $\hat{u}_{\text{rms}}$ and $\hat{a}_z$ with $\tau < 0$. The red region corresponds to significantly large fluctuation that require a strong bank, whereas cues in the blue regions are filtered out. The acceleration $\hat{a}_z$ is normalized by $v_{\text{glider}}/\Delta t$, where $\Delta t = 1$ s and v_{glider} is the speed of the glider.

$$\mu_{\text{opt}}(\boldsymbol{a}_z,\tau)=\arg\max_{\mu^{t+1}}\Pr\left(\mu^{t+1}|\boldsymbol{a}_z^t,\tau^t\right), \quad [8]$$

and we are interested in the variations of the optimal bank angle with the turbulence level $\hat{u}_{\text{rms}}$. We use a bicubic spline interpolation to smooth the probability distributions and thereby obtain smoothened values for μ_{opt}.

To create a higher resolution in $\boldsymbol{a}_z$, we expand our state space by creating finer divisions in the vertical wind accelerations. Note that the performance with an expanded state space is not significantly better than the one with just three states. Fig. 4 shows a heat map of the optimal bank angles at different $\boldsymbol{a}_z<0$ and $\tau<0$. For every $\boldsymbol{a}_z$, μ_{opt} drops from the maximum value of 15° to a value closer to zero as $\hat{u}_{\text{rms}}$ increases. Note that, because $\tau<0$, the optimal angles are biased toward being positive. We define a threshold on the optimal bank angles at 12.5°, which empirically corresponds to the point where the optimal bank angles drop most rapidly as $\hat{u}_{\text{rms}}$ increases. Above (below) the threshold, the angles are considered "high" ("low"). The threshold on the optimal bank angle defined a cutoff on $-\boldsymbol{a}_z$ and thereby an effective "fluctuation filter."

We interpret the fluctuation filter above as follows: at a particular $\hat{u}_{\text{rms}}$, if the glider encounters a fluctuation with $-\boldsymbol{a}_z$ above the cutoff, the glider interprets the fluctuation as significant, that is, as the large-scale downward branch of a convective cell, and banks away. Conversely, fluctuations below the cutoff are ignored. In other words, the cutoff defined above gives the level that identifies significantly large fluctuations that require action. Similar behaviors are obtained for $(\boldsymbol{a}_z<0,\tau=0)$ and $\tau>0$ is symmetric with respect to the case $\tau<0$ just discussed. Conversely, for $\boldsymbol{a}_z>0$, the glider maintains a bank angle close to zero unless it experiences an exceptionally large torque. These simple principles are the key for effective soaring in fluctuating turbulent environments.

Discussion

We have shown that reinforcement learning methods cope with strong turbulent fluctuations and identify effective policies of navigation in turbulent flow. Previous works neglected turbulence, which is an essential and unavoidable feature of natural flow. The learned policies dramatically improve the gain of height and the rapidity of climbing within thermals, even when turbulent fluctuations are strong and the glider has reduced control due to its being transported by the flow.

We deliberately kept simple the sensorimotor cues that the glider can sense to guide its flight. In particular, possible cues were local in space and time for two reasons: keep the closest contact with what birds are likely to sense and minimize the mechanical instrumentation needed for the control of autonomously flying vehicles. In the same spirit, we kept simple the parametrization of the learned policies, by using a relatively coarse discretization of the space of states and actions.

Turbulence has indeed a major impact upon the policy of flight. We explicitly presented how the learned policies of flight modify as the level of turbulence increases. In particular, we quantified the increase of the threshold on the cues needed for the glider to change its parameters of control. We also discussed the simple principles that the policy follows to filter out transient, small-scale turbulent fluctuations, and identify the level of the sensorimotor cues that requires actions that modify the parameters of flight of the glider.

We found that the bank angle of the glider is the main control for navigation within a single thermal, which is the main interest of the current work. However, we also considered a very simplified setting mimicking the flight between multiple thermals, and there we found that control of the angle of attack is important. Interthermals flight is of major interest for birds' migration and glider pilots. MacCready (28) determined the optimal speed to maximize the average cross-country speed as a function of the glider's rate of sink and the velocity of ascent within the thermals. The resulting instrument (the so-called MacCready speed ring) is commonly used by glider pilots with various supplementary empirical prescriptions, which typically tend to be risk averse. MacCready's prediction was also recently compared with the behavior of various birds (29) along their thermal-dense migratory routes. Their behavior was found to differ from the prediction; namely, a more conservative policy was observed, with slower but less sinking paths that reduce the probability of dramatic losses of height. One possible cause for more conservative policies relates to the uncertainties on the location and the velocity of ascent within the thermals, which was previously considered in the literature (30). Another possible reason suggested by our results is turbulence along the interthermal paths, which is neglected in MacCready's and subsequent arguments. Our methodology can be adapted to realistically model interthermal conditions, and future work will assess the role of turbulence in the policy of interthermal flight.

We identified torque and vertical accelerations as the local sensorimotor cues that most effectively guide turbulent navigation. Temperature was specifically shown to yield minor gains. The robustness of our results with respect to the modeling of turbulence strongly suggests that the conclusion applies to natural conditions; a sensor of temperature could then be safely spared in the instrumentation for autonomous flying vehicles. More generally, it will be of major interest to implement our predicted policy on remotely controlled gliders and test their flight performance in field experiments. Thanks to our choices discussed above, the mechanical instrumentation needed for control is minimal and can be hosted on commercial gliders without perturbing their aerodynamics. Finally, our flight policy and the nature of the sensorimotor cues that we identified, provide predictions that can be compared with the behavior of soaring birds and could shed light on the decision processes that enable them to perform their soaring feats.

Methods

Our kinematic model of turbulence extends the one in ref. 21 to the inhomogeneous case relevant for the atmospheric boundary layer. We can thereby statistically reproduce the Kolmogorov and Richardson laws (10) and the velocity profile of the atmospheric boundary layer (6). The atmospheric boundary layer on a sunny day extends to an inversion height $z_i\sim 1$ km and mainly consists of two layers—the free convection layer, extending up to $0.1z_i$, and the mixed layer (6). The rms of velocity fluctuations varies with the height z as $\langle\delta(u^{\text{kin}})^2(z)\rangle\sim z^{2/3}$ in the free convection layer and is statistically constant in the mixed layer. To reproduce these statistics, we decomposed the velocity field at height z into contributions from fields of different integral length scales l_n:

$$\boldsymbol{u}^{\text{kin}}(\boldsymbol{x}_\perp,z,t)=\sum_{l_n>z}c_n\boldsymbol{u}^{\text{kin}}(\boldsymbol{x}_\perp,z,t|l_n), \quad [9]$$

where $\boldsymbol{x}_\perp$ are the two horizontal components of the position. The velocity field at each length scale ℓ_n is specified in spatial wavenumbers $\boldsymbol{k}$ as follows:

$$\boldsymbol{u}^{\text{kin}}(\boldsymbol{x}_\perp,z,t|l_n)=\int\hat{\boldsymbol{u}}_n^{\text{kin}}(\boldsymbol{k},t)e^{i\boldsymbol{k}.\boldsymbol{x}}\mathrm{d}^3\boldsymbol{k}, \quad [10]$$

where the individual Fourier components $\hat{\boldsymbol{u}}_n^{\text{kin}}(\boldsymbol{k},t)$ are modeled as a Ornstein–Uhlenbeck process (21). The corresponding diffusion constant is set such that the spatial energy spectrum follows the Kolmogorov five-thirds law $E(k)\sim k^{-5/3}$, where $k=|\boldsymbol{k}|$. The power law energy spectrum gives rise to long-range spatial correlations with fluctuations at every length scale up to l_n. The relaxation time of each mode is given by the Kolmogorov scaling $\tau_k\sim k^{-2/3}$ (10). The coefficients c_n and the integral length scales l_n are chosen to reproduce the velocity profile of the boundary layer (see *Supporting Information* for details). We accounted for the mean ascending current within the thermals by superposing a Gaussian-shaped mean vertical velocity on top of the fluctuations.

ACKNOWLEDGMENTS. We are grateful to A. Libchaber for numerous discussions on convective flow. This work was supported by Simons Foundation Grant 340106 (to M.V.).

1. Cone CD, Jr (1962) Thermal soaring of birds. *Am Sci* 50(1):180–209.
2. Pennycuick CJ (1983) Thermal soaring compared in three dissimilar tropical bird species, *Fregata magnificens*, *Pelecanus occidentalis* and *Coragyps atratus*. *J Exp Biol* 102:307–325.
3. Ehrlich P, Dobkin D, Wheye D (1988) *The Birder's Handbook: A Field Guide to the Natural History of North American Birds*. (Simon and Schuster, New York).
4. Newton I (2007) *The Migration Ecology of Birds* (Academic, London).
5. Allen MJ (2007) Guidance and control of an autonomous soaring UAV. *Proceedings of the AIAA Aerospace Sciences Meeting* (American Institute of Aeronautics and Astronautics, Reston, VA), AIAA Paper 2007-867.
6. Garrat JR (1994) *The Atmospheric Boundary Layer*. Cambridge Atmospheric and Space Science Series (Cambridge Univ Press, Cambridge, UK).
7. Lenschow DH, Stephens PL (1980) The role of thermals in the convective boundary layer. *Boundary-Layer Meteorol* 19(4):509–532.
8. Young GS (1988) Convection in the atmospheric boundary layer. *Earth Sci Rev* 25(3): 179–198.
9. Ahlers G, Grossmann S, Lohse D (2009) Heat transfer and large scale dynamics in turbulent Rayleigh–Benard convection. *Rev Mod Phys* 81:503.
10. Frisch U (1995) *Turbulence: The Legacy of A. N. Kolmogorov* (Cambridge Univ Press, Cambridge, UK).
11. Akos Z, Nagy M, Vicsek T (2008) Comparing bird and human soaring strategies. *Proc Natl Acad Sci USA* 105(11):4139–4143.
12. Sutton RS, Barto AG (1998) *Reinforcement Learning: An Introduction* (MIT Press, Cambridge, MA).
13. Wharington J, Herszberg I (1998) Control of a high endurance unmanned aerial vehicle. *Proceedings of the 21st Congress of International Council of the Aeronautical Sciences* (International Council of the Aeronautical Sciences, Bonn, Germany), Paper 98-3.7.1.
14. Reichmann H (1988) *Cross-Country Soaring* (Thomson Publications, Santa Monica, CA).
15. Lawrance NRJ, et al. (2014) Long endurance autonomous flight for unmanned aerial vehicles. *AerospaceLab* 8(5):1–15.
16. Woodbury T, Dunn C, Valasek J (2014) Autonomous soaring using reinforcement learning for trajectory generation. *Proceedings of the AIAA Aerospace Sciences Meeting* (American Institute of Aeronautics and Astronautics, Reston, VA), AIAA Paper 2014-0990.
17. Akos Z, Nagy M, Leven S, Vicsek T (2010) Thermal soaring flight of birds and unmanned aerial vehicles. *Bioinspir Biomim* 5(4):045003.
18. Popinet S (2003) Gerris: A tree-based adaptive solver for the incompressible Euler equations in complex geometries. *J Comput Phys* 190:572–600.
19. Verzicco R, Camussi R (2003) Numerical experiments on strongly turbulent thermal convection in a slender cylindrical cell. *J Fluid Mech* 477:19–49.
20. Verzicco R, Camussi R (1999) Prandtl number effects in convective turbulence. *J Fluid Mech* 383:55–73.
21. Fung JCH, Hunt JCR, Malik NA, Perkins RJ (1992) Kinematic simulation of homogeneous turbulence by unsteady random Fourier modes. *J Fluid Mech* 236:281–318.
22. von Mises R (1945) *Theory of Flight* (McGraw Hill, New York).
23. Anderson JR, Jr (1978) *Introduction to Flight* (McGraw Hill, New York).
24. von Karman T (1963) *Aerodynamics* (McGraw-Hill, New York).
25. Tesauro G (1995) Temporal difference learning and TD-Gammon. *Commun ACM* 38: 58–68.
26. Shraiman BI, Siggia ED (2000) Scalar turbulence. *Nature* 405(6787):639–646.
27. Falkovich G, Gawedzki K, Vergassola M (2001) Particles and fields in fluid turbulence. *Rev Mod Phys* 73:913–975.
28. MacCready PBJ (1958) Optimum airspeed selector. *Soaring* 1958(1):10–11.
29. Horvitz N, et al. (2014) The gliding speed of migrating birds: Slow and safe or fast and risky? *Ecol Lett* 17(6):670–679.
30. Cochrane JH (1999) MacCready theory with uncertain lift and limited altitude. *Technical Soaring* 23:88–96.
31. Grotzbach G (1983) Spatial resolution requirements for direct numerical simulation of the Rayleigh–Bénard convection. *J Comput Phys* 49:241–264.