LETTER

https://doi.org/10.1038/s41586-018-0533-0

Glider soaring via reinforcement learning in the field

Gautam Reddy[1,5], Jerome Wong-Ng[1,5], Antonio Celani[2], Terrence J. Sejnowski[3,4] & Massimo Vergassola[1]*

Soaring birds often rely on ascending thermal plumes (thermals) in the atmosphere as they search for prey or migrate across large distances[1–4]. The landscape of convective currents is rugged and shifts on timescales of a few minutes as thermals constantly form, disintegrate or are transported away by the wind[5,6]. How soaring birds find and navigate thermals within this complex landscape is unknown. Reinforcement learning[7] provides an appropriate framework in which to identify an effective navigational strategy as a sequence of decisions made in response to environmental cues. Here we use reinforcement learning to train a glider in the field to navigate atmospheric thermals autonomously. We equipped a glider of two-metre wingspan with a flight controller that precisely controlled the bank angle and pitch, modulating these at intervals with the aim of gaining as much lift as possible. A navigational strategy was determined solely from the glider's pooled experiences, collected over several days in the field. The strategy relies on on-board methods to accurately estimate the local vertical wind accelerations and the roll-wise torques on the glider, which serve as navigational cues. We establish the validity of our learned flight policy through field experiments, numerical simulations and estimates of the noise in measurements caused by atmospheric turbulence. Our results highlight the role of vertical wind accelerations and roll-wise torques as effective mechanosensory cues for soaring birds and provide a navigational strategy that is directly applicable to the development of autonomous soaring vehicles.

In reinforcement learning, an animal maximizes its long-term reward by taking actions in response to its external environment and internal state. Learning occurs by reinforcing behaviour based on feedback from past experiences. Similar ideas have been used to develop intelligent agents that have achieved spectacular performance in strategic games such as backgammon[8] and Go[9], visual-based video game play[10] and robotics[11,12]. In the field, however, constraints imposed by variable and uncontrolled conditions prevent learning agents from using data-intensive learning algorithms and the optimization of model design needed for quicker learning. These are the conditions most often faced by living organisms.

A striking example in nature is provided by thermal soaring. Atmospheric convection is not consistent across days and, even under suitable conditions, the locations, sizes, durations and strengths of nearby thermals are unpredictable. As a result, the distribution of samples used to train the glider differs day-to-day. Gliders and birds operate at spatial and temporal scales where fluctuations in wind velocities are due to turbulent eddies lasting a few seconds that may mask or falsely enhance a glider's estimate of its mean climb rate. Further, the measurement of navigational cues using standard instrumentation may be consistently biased by aerodynamic effects that require precise quantification. Here, we demonstrate that reinforcement learning can meet the challenge of learning to soar effectively in atmospheric turbulent environments. In past work, by contrast, the manoeuvring of an autonomous helicopter in ref. [11] is a control problem that is decoupled from environmental fluctuations and has little trial-to-trial variability. Past autonomous soaring algorithms have largely relied on locating the centroid of a drifting Gaussian thermal[13–16], which is unrealistic, or have applied learning methods in highly simplified simulated settings[17–19].

Using the reinforcement learning framework[7], we may describe the behaviour of the glider as an agent traversing different states (s) by taking actions (a) while receiving a local reward (r). The goal is to find a behavioural policy that maximizes the 'value': that is, the mean sum of future rewards up to a specified horizon. We seek a model-free approach, which estimates the value of different actions at a particular state (called the Q function) solely through the agent's experiences during repeated instances of the task, thereby bypassing the modelling of complex atmospheric physics and aerodynamics (see Methods). The optimal policy is subsequently derived by taking actions with the highest Q value at each state, where the state includes sensorimotor cues and the glider's aerodynamic state.

To identify mechanosensory cues that could guide soaring, we recently combined the above ideas with simulations of virtual gliders in numerically generated turbulent flow[20]. Two cues emerged from our screening: (1) the vertical wind acceleration (a_z) along the glider's path and (2) the spatial gradients in the vertical wind velocity across the wings of the glider (ω). Intuitively, the two cues correspond to the gradient of the vertical wind velocity in the longitudinal and lateral directions of the glider, which locally orient it towards regions of higher lift. Simulations[20] further showed that the glider's bank angle is the crucial aerodynamic control variable; additional variables such as the angle of attack, or other mechanosensory cues such as temperature or vertical velocity, offer minor improvements when navigating within a thermal.

To learn to soar in the field, a glider (wingspan, 2 m) was equipped with autonomous soaring capabilities (Fig. 1a, b). The glider is equipped with a flight controller, which uses a feedback control system to modulate the glider's ailerons and elevator such that the bank angle and pitch take the values desired by the behavioural policy being used (we use two different behavioural policies during initial learning, and the gliders then implement a further policy—the final navigational strategy—after learning). Relevant measurements, such as the altitude, ground velocity ($\boldsymbol{u}$), airspeed, bank angle (μ) and pitch, are made continuously at 10 Hz with standard instrumentation (see Methods). At fixed time intervals, the glider changes its heading by modulating its bank angle in accordance with the implemented behavioural policy.

Noise and biases that affect learning in the field require the development of appropriate methods to extract environmental cues from measurements made by sensory devices. We found that estimates of a_z from the derivative of the vertical ground velocity (u_z) are biased by longitudinal motions of the glider about the pitch axis as the glider responds to an imbalance of forces and moments while turning. By modelling the glider's longitudinal dynamics, we obtain an unbiased estimate of the local vertical wind velocity (w_z), and a_z as its derivative (see Methods). The estimation of the spatial gradients across the wings, ω, poses a greater challenge, as it involves the difference between two noisy measurements at relatively close positions. The key observation that we used here is that the glider rolls because of contributions from vertical wind velocity gradients, the feedback control mechanism and various aerodynamic effects. The resulting roll-wise torque can be estimated from the small deviations of the true bank angle from

[1]Department of Physics, University of California, San Diego, La Jolla, CA, USA. [2]The Abdus Salam International Center for Theoretical Physics, Trieste, Italy. [3]The Salk Institute for Biological Studies, La Jolla, CA, USA. [4]Division of Biological Sciences, University of California, San Diego, La Jolla, CA, USA. [5]These authors contributed equally: Gautam Reddy, Jerome Wong-Ng.
*e-mail: massimo@physics.ucsd.edu

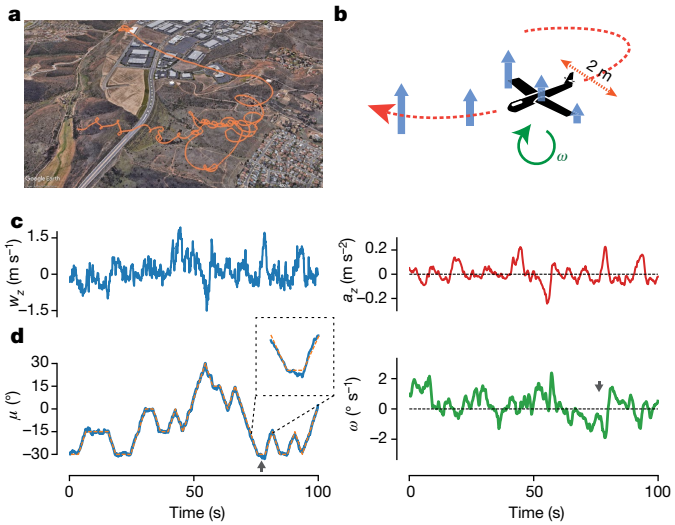


Fig. 1 | Soaring in the field by using turbulent navigational cues. **a**, A trajectory (orange line) of our glider soaring in Poway, California. **b**, A cartoon of the glider showing the available navigational cues—gradients in vertical wind velocities (indicated by the length of the blue arrows) along the trajectory and across its wings, which generate a vertical wind acceleration a_z and a roll-wise torque ω, respectively. **c**, A sample trace of the estimated vertical wind velocity w_z and corresponding a_z obtained in the field. **d**, The measured bank angle μ and the estimated ω during the same trial as in **c**. ω (solid green line) is estimated from the small deviations of the measured bank angle (solid blue line) from the expected bank angle (dashed orange line) after accounting for other effects (see Methods). The black arrows mark the enlarged bank angle trajectory shown in the inset in the left panel.

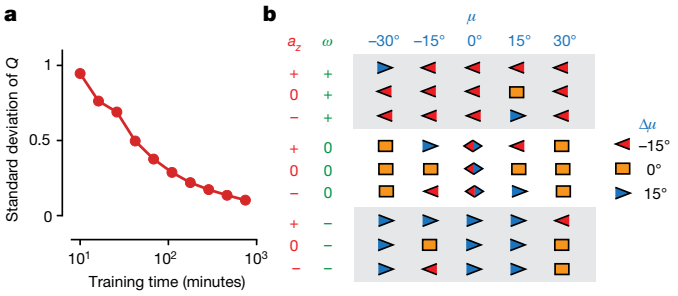


Fig. 2 | Convergence of the learning algorithm and the learned strategy for navigating thermal plumes. **a**, The convergence of Q values during learning as measured by the standard deviation of the mean Q value versus training time in the field, obtained by bootstrapping from the experiences accumulated up to that point. **b**, The final learned policy. Each symbol corresponds to the best action (increasing or decreasing the bank angle μ by 15° or maintaining the same μ, as shown in the key on the right) to be taken when the glider observes a particular (a_z, ω) pair and is banked at μ. Combined symbols depict pairs of actions that are equally rewarding. A positive (negative) ω corresponds to a higher vertical wind velocity on the left (right) wing of the glider and a positive (negative) μ corresponds to turning right (left) with respect to the glider's heading.

the desired one, and a new dynamical model allows us to separate the ω contribution due to velocity gradients from the other effects (see Methods). A sample trace of the resulting unbiased estimate of ω is shown in Fig. 1c, d, together with traces of w_z, μ and unbiased estimates of a_z.

Equipped with a proper procedure for estimating environmental cues, we next addressed the specifics of learning in the field. First, to constrain our state space, we discretized the range of values of a_z and ω into three states each: positive high (+), neutral (0) and negative high (−). Second, we found that learning is accelerated by choosing a_z attained at the subsequent time step as the reward signal. The choice of a_z (rather than w_z) is an instance of reward shaping that is justified in Supplementary Information, where we show that using a_z as a reward still leads to a policy that optimizes the long-term gain in height. This property is a special case of our general result that a particular reward function or its time derivatives (of any order) yield the same optimal policy (Supplementary Information). Choosing w_z as the reward fails to drive learning in the soaring problem, possibly because the velocities (and thus the rewards) are correlated across states and their temporal statistics strongly deviates from the Markovianity assumption in reinforcement learning methods[7]. Velocity fluctuations in turbulent flow are long-correlated: that is, their correlation timescale is determined by the largest timescale of the flow (see, for instance, figure 9 of ref. [21]), which is of the order of minutes in the atmosphere. Conversely, the correlation timescale of accelerations is controlled by the smallest timescale[21–23] (the dissipation timescale in figure 7 of ref. [21]). This is estimated to be only a fraction of a second, which is much smaller than the time interval between successive actions. Note that the previous experimental observations can be rationalized by the combination of the power-law spectrum of turbulent velocity fluctuations in the atmosphere and the extra factor of frequency squared in the spectrum of acceleration versus velocity fluctuations[23]. Finally, the glider's experiences, represented as state–action–state–reward quadruplets, (s_t, a_t, s_{t+1}, r_t), were cumulatively collected (over 15 days) into a set E using explorative behavioural policies. Learning is monitored by bootstrapping the standard deviation of the Q values from E (Fig. 2a), calculated through value iteration methods (see Methods).

The navigational strategy derived at the end of the training period is presented in Fig. 2b, which shows the actions deemed optimal for the 45 possible states. The rows corresponding to $\omega = 0$ resemble the Reichmann rules[24]—a set of simple heuristics for soaring, which suggest a decrease (increase) in bank angle when the climb rate increases (decreases). Our strategy also gives a prescription for bank: for instance, when a_z and ω are both positive (top row in Fig. 2b)—that is, in a situation when better lift is available diagonal to the glider's heading—it is advantageous not to bank to the extreme but rather to maintain an intermediate value between −30° and −15°. Importantly, the learned leftward (rightward) bias in bank angle on encountering a positive (negative) torque validates our estimation procedure for ω.

In Fig. 3a, we show a sample trajectory of a glider that used the navigational strategy in the field to remain aloft for about 12 min while spiralling to the height of low-lying clouds (see also Extended Data Fig. 1). On a day with strong atmospheric convection, the time spent aloft is limited only by visibility and the receiver's range as the glider soars higher or is constantly pushed away by the wind. A significant improvement in median climb rate of 0.35 m s^{-1} was measured in the field by performing repeated 3-min trials over 5 days (Fig. 3b, Mann–Whitney $U = 429$, $n_{\text{control}} = 37$, $n_{\text{strategy}} = 49$, $P < 10^{-4}$ two-sided). Notably, this value reflects a general improvement in performance averaged across widely variable conditions without controlling for the availability of nearby thermals.

To examine possible advantages of larger gliders due to improved estimation of torque, we further analysed soaring performance for different wingspans (l). Although the naive expectation is that the signal-to-noise ratio in the estimate of ω scales linearly with l, we show that the effects of atmospheric turbulence lead to a much weaker $l^{1/6}$ scaling (see Methods). Because testing our prediction would require a series of gliders with different wingspans, we turned to numerical simulations of the convective boundary layer, adapted to reflect our experimental set-up (Methods). Results shown in Fig. 3c, d are consistent with the predicted scaling. Intuitively, the weak 1/6 exponent arises because the improvement in estimation of the gradient is offset by the larger turbulent eddies, which only have a sweeping effect for smaller wingspans (that is, they do not rotate the glider but translate it, which does not affect the estimate of vertical velocity differences across its wings), and contribute to velocity differences across the wings as l increases. Our calculation yields an estimate of the signal-to-noise ratio of about 4 for typical experimental values; similar arguments for a_z yield a signal-to-noise ratio of about 7. Experimental results, together with

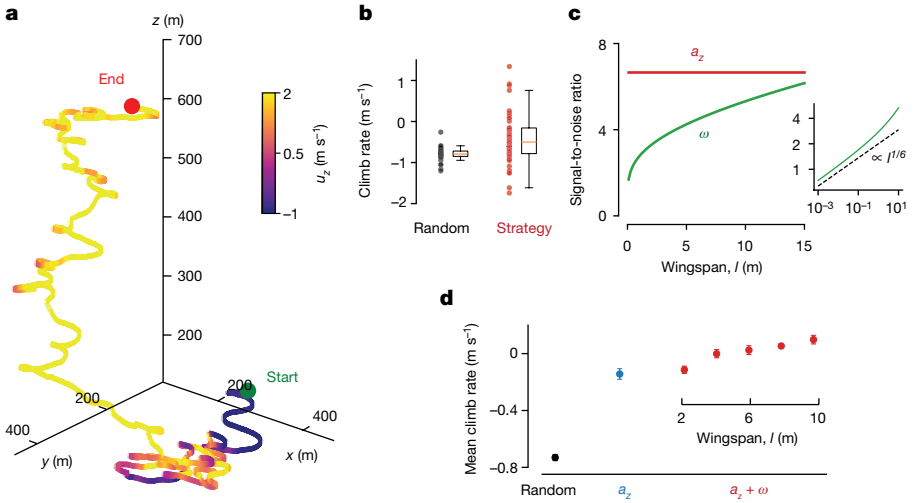


Fig. 3 | Performance of the learned strategy and its dependence on the wingspan. a, A 12-min-long trajectory of the glider executing the learned strategy for navigating thermals in the field, coloured according to the vertical ground velocity at each instant. **b**, Experimentally measured climb rate for a control random policy (black dots) is compared against the learned strategy (red dots) over repeated 3-min trials in the field. Each dot represents the average climb rate in a single trial. To restrict the range of the axis, a few outliers are not shown. The limits on the y axis are from -2 m s^{-1} to 1.5 m s^{-1}. The orange line in the box plot shows the median, the extent of the boxes marks the interquartile range and the whiskers demarcate the outliers (1.5 times the interquartile range above (below) the upper (lower) quartile). **c**, Signal-to-noise ratio for estimating ω (green) and a_z (red) as a function of wingspan. The signal-to-noise ratio for ω estimation is plotted in logarithmic scale (inset) to highlight the weak $l^{1/6}$ scaling. **d**, The mean climb rate for the learned strategy is compared for different wingspans (red filled circles) in simulations of a glider soaring in the convective boundary layer. For comparison, we show the mean climb rates for a random policy and for a strategy that uses a_z only (see Methods). Error bars represent s.e.m.

simulations and signal-to-noise ratio estimates, establish a_z and ω as robust navigational cues for thermal soaring.

The real-world intricacies of soaring impose severe constraints on the complexity of the underlying models, reflecting a fundamental trade-off between learning speed and performance. Notably, the choice of a proper reward signal was crucial to make learning feasible with the limited samples available. Although reward shaping has received some attention in the machine learning community[25], its relevance to animal behaviour remains poorly understood. We remark that our navigational strategy constitutes a set of general reactive rules, with no learning occurring during a particular thermal encounter. A soaring bird may use a model-based approach of constantly updating its estimate of the location of nearby thermals based on recent experience and visual cues. Still, the importance of vertical wind accelerations and torques for our policy suggests that they are likely to be useful for any other strategy; our methods of estimating them in a glider suggest that they should be accessible to birds as well. The hypothesis that birds use those mechanical cues while soaring can be tested in experiments.

Finally, we note that single-thermal soaring is just one face of a multifaceted question: how should a migrating bird or a cross-country glider fly among thermals over hundreds of kilometres for a quick, yet risk-averse, journey[26–28]? This calls for the development of effective methods for identifying areas of strong updraft based on mechanical and visual cues. Such methods, coupled with our current work, would pave the way to a better understanding of how birds migrate and the development of autonomous vehicles that can fly for long distances and long periods with minimal energy cost.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0533-0.

Received: 20 February 2018; Accepted: 20 July 2018;
Published online 19 September 2018.

1. Newton, I. *Migration Ecology of Soaring Birds* 1st edn (Elsevier, Amsterdam, 2008).
2. Shamoun-Baranes, J., Leshem, Y., Yom-tov, Y. & Liechti, O. Differential use of thermal convection by soaring birds over central Israel. *Condor* **105**, 208–218 (2003).
3. Weimerskirch, H., Bishop, C., Jeanniard-du-Dot, T., Prudor, A. & Sachs, G. Frigate birds track atmospheric conditions over months-long transoceanic flights. *Science* **353**, 74–78 (2016).
4. Pennycuick, C. J. Thermal soaring compared in three dissimilar tropical bird species, *Fregata magnificens*, *Pelecanus occidentals* and *Coragyps atratus*. *J. Exp. Biol.* **102**, 307–325 (1983).
5. Garrat, J. R. *The Atmospheric Boundary Layer* (Cambridge Univ. Press, Cambridge, 1994).
6. Lenschow, D. H. & Stephens, P. L. The role of thermals in the atmospheric boundary layer. *Boundary-Layer Meteorol.* **19**, 509–532 (1980).
7. Sutton, R. S. & Barto, A. G. *Reinforcement Learning: An Introduction* 1st edn (MIT Press, Cambridge, 1998).
8. Tesauro, G. Temporal difference learning and TD-Gammon. *Commun. ACM* **38**, 58–68 (1995).
9. Silver, D. et al. Mastering the game of Go without human knowledge. *Nature* **550**, 354–359 (2017).
10. Mnih, V. et al. Human-level control through deep reinforcement learning. *Nature* **518**, 529–533 (2015).
11. Kim, H. J., Jordan, M. I., Sastry, S. & Ng, A. in *Advances in Neural Information Processing Systems* Vol. 16 (eds Thrun, S. et al.) 799–806 (MIT Press, Cambridge, 2004).
12. Levine, S., Finn, C., Darrell, T. & Abbeel, P. End-to-end training of deep visuomotor policies. *J. Mach. Learn. Res.* **17**, 1–40 (2016).
13. Allen, M. J. & Lin, V. Guidance and control of an autonomous soaring vehicle with flight test results. In *45th AIAA Aerospace Sciences Meeting and Exhibit* 2007-867 (AIAA, 2007).
14. Edwards, D. J. Implementation details and flight test results of an autonomous soaring controller. In *AIAA Guidance, Navigation and Control Conference and Exhibit* 2008-7244 (AIAA, 2008).
15. Edwards, D. J. *Autonomous Soaring: The Montague Cross Country Challenge*. PhD thesis, North Carolina State Univ. (2010).
16. Ákos, Z., Nagy, M., Leven, S. & Vicsek, T. Thermal soaring flight of birds and unmanned aerial vehicles. *Bioinspir. Biomim.* **5**, 045003 (2010).
17. Doncieux, S., Mouret, J. B. & Meyer, J.-A. Soaring behaviors in UAVs: 'animat' design methodology and current results. In *3rd US–European Competition and Workshop on Micro Air Vehicle Systems (MAV07) and European Micro Air Vehicle Conference and Flight Competition (EMAV2007)* (2007); http://www.isir.upmc.fr/files/2007ACTI734.pdf.
18. Wharington, J. & Herszberg, I. Control of a high endurance unmanned aerial vehicle. In *21st Congress of International Council of the Aeronautical Sciences* 98-3.7.1 (ICAS, 1998).
19. Chung, J. J., Lawrance, N. R. J. & Sukkarieh, S. Learning to soar: resource-constrained exploration in reinforcement learning. *Int. J. Robot. Res.* **34**, 158–172 (2015).
20. Reddy, G., Celani, A., Sejnowski, T. & Vergassola, M. Learning to soar in turbulent environments. *Proc. Natl Acad. Sci. USA* **113**, E4877–E4884 (2016).
21. Yeung, P. K. & Pope, S. B. Lagrangian statistics from direct numerical simulations of isotropic turbulence. *J. Fluid Mech.* **207**, 531–586 (1989).

22. Voth, G. A., La Porta, A., Crawford, A. M., Alexander, J. & Bodenschatz, E. Measurement of particle accelerations in fully developed turbulence. *J. Fluid Mech.* **469**, 121–160 (2002).
23. Tennekes, H. & Lumley, J. L. *A First Course in Turbulence* (MIT Press, Cambridge, 1972).
24. Reichmann, H. *Cross-Country Soaring* (Thomson Publications, Santa Monica, 1988).
25. Ng, A. Y., Harada, D. & Russell, S. J. Policy invariance under reward transformations: theory and application to reward shaping. In *Proc. 16th International Conference on Machine Learning* (eds Bratko, I. & Dzeroski, S.) 278–287 (Morgan Kaufmann, San Francisco, 1999).
26. MacCready, P. B. J. Optimum airspeed selector. *Soaring* **1958**, 10–11 (1958).
27. Horvitz, N. et al. The gliding speed of migrating birds: slow and safe or fast and risky? *Ecol. Lett.* **17**, 670–679 (2014).
28. Cochrane, J. H. MacCready theory with uncertain lift and limited altitude. *Tech. Soaring* **23**, 88–96 (1999).

Acknowledgements This work was supported by Simons Foundation grant 340106 (to M.V.) and NSF grant NCS-FO-1735004 (to T.J.S.).

Reviewer information *Nature* thanks M. Chertkov and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions All authors were involved in designing the study and drafting the final manuscript. G.R. and J.W.N. performed the experiments and analysed the data. G.R., A.C. and M.V. contributed to the theoretical results.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0533-0.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0533-0.

Correspondence and requests for materials should be addressed to M.V.

METHODS

Experimental set-up. A Parkzone Radian Pro fixed-wing plane of 2-m wingspan was equipped with an on-board Pixfalcon autonomous flight controller operating on custom-modified Arduplane firmware (http://www.ardupilot.org). The instrumentation available to the flight controller includes a GPS, compass, barometer, airspeed sensor and an inertial measurement unit. Measurements from multiple instruments are combined by an extended Kalman filter (EKF) to give an estimate of relevant quantities such as the altitude z, the sink rate with respect to the ground $-u_z$, pitch ϕ, bank angle μ and the airspeed V, at a rate of 10 Hz (see Extended Data Fig. 2 for the definitions of the angles). Throughout the paper, we use $\mu > 0$ when the glider is banked to the right and $\phi > 0$ for the glider pitched with its nose above the horizontal plane. For a given desired pitch ϕ_d and desired bank angle μ_d, the controller modulates the aileron and elevator control surfaces at 400 Hz by using a proportional–integral–derivative feedback control mechanism at a user-set timescale τ (see Extended Data Table 1 for parameter values) such that:

$$\tau \frac{d\phi}{dt} = \phi_d - \phi \tag{1}$$

$$\tau \frac{d\mu}{dt} = \mu_d - \mu \tag{2}$$

The desired pitch is fixed during flight and can be used to indirectly modulate the angle of attack, α, which determines the airspeed and sink rate with respect to air of the glider ($-v_z$). Actions of increasing, decreasing or keeping the same bank angle are taken in time steps of t_a by changing μ_d such that μ increases linearly from μ_i to μ_f in time interval t_a:

$$\mu_d(t) = \mu_i + (\mu_f - \mu_i)\frac{t+\tau}{t_a} \tag{3}$$

Estimation of the vertical wind acceleration. The vertical wind acceleration is defined as:

$$a_z \equiv \frac{dw_z}{dt} = \frac{d}{dt}(u_z - v_z) \tag{4}$$

where $\boldsymbol{u}$ and $\boldsymbol{v}$ are the velocities of the glider with respect to the ground and air respectively, and $\boldsymbol{w}$ is the wind velocity. Here, we have used the relation $\boldsymbol{w} = \boldsymbol{u} - \boldsymbol{v}$. An estimate of $\boldsymbol{u}$ is obtained in a straightforward manner from the EKF, which combines the GPS and barometer readings to form the estimate. However, v_z is confounded by various aerodynamic effects that affect it on timescales of a few seconds (Extended Data Fig. 3). Artificial accelerations introduced by these effects impair accurate estimation of the wind acceleration and thus alter the perceived state during decision-making and learning. Two effects strongly affect variations in v_z: (1) sustained pitch oscillations with a period of a few seconds and varying amplitude, and (2) variations in angle of attack, which occur to compensate for the imbalance of lift and weight while rolling. In Supplementary Information, we present a detailed analysis of the longitudinal motions that affect the glider, summarized here for conciseness. Changes in v_z can be approximated as:

$$\Delta v_z = -V(\Delta\alpha - \Delta\phi) \tag{5}$$

where Δ denotes the deviation from their value during steady, level flight. We obtain $\Delta\phi$ directly from on-board measurements, whereas $\Delta\alpha$ can be approximated for bank angle μ as:

$$\Delta\alpha \approx (\alpha_0 - \alpha_i)\left(\frac{1}{\cos\mu} - 1\right) \tag{6}$$

where α_0 is the angle of attack at steady, level flight and α_i is a parameter that depends on the geometry and the angle of incidence of the wing. The constant pre-factor $(\alpha_0 - \alpha_i)$ is inferred from experiments. Measurements of u_z together with the estimate of Δv_z are now used to estimate the vertical wind velocity w_z up to a constant term, which can be ignored as it does not affect a_z. The vertical wind acceleration a_z is then obtained by taking the derivative of w_z and is further smoothed using an exponential smoothing kernel of timescale σ_a (Extended Data Fig. 4).

Estimation of vertical wind velocity gradients across the wings. Spatial gradients in the vertical wind velocity induce a roll-wise torque on the plane, which we estimate using the deviation of the measured bank angle from the expected bank angle. The total roll-wise torque on the plane has contributions from three sources: (1) the feedback control of the plane; (2) spatial gradients in the wind including turbulent fluctuations; and (3) roll-wise moments due to various aerodynamic effects. Here, we follow an empirical approach: we note that the latter two contributions perturb the evolution of the bank angle from equation (2). We can then write an effective equation

$$\frac{d\mu}{dt} = \frac{\mu_d - \mu}{\tau} + \omega(t) + \omega_{aero}(t) \tag{7}$$

where $\omega(t)$ and $\omega_{aero}(t)$ are contributions to the roll-wise angular velocity due to the wind and aerodynamic effects, respectively. We empirically find four major contributions to ω_{aero}: (1) the dihedral effect, which is a stabilizing moment due to the effects of sideslip on a dihedral wing geometry; (2) the overbanking effect, which is a destabilizing moment that occurs during turns with small radii; (3) trim effects, which create a constant moment due to asymmetric lift on the two wings; and (4) a loss of rolling moment generated by the ailerons when rolling at low airspeeds. We quantify the contributions from the four effects and model their dependence on the bank angle (see Supplementary Information for more details on modelling and calibration). An estimate of ω is then obtained as:

$$\omega = \frac{d\mu}{dt} - \frac{\mu_d - \mu}{\tau} - \omega_{aero} \tag{8}$$

Finally, an exponential smoothing kernel is applied to obtain a smoothed ω (Extended Data Fig. 5).

Design of the learning module. The navigational component of the glider is modelled as a Markov decision process, closely following the implementation used in ref. [20]. The Markovian transitions are discretized in time into intervals of size t_a. The state space consists of the possible values taken by a_z, ω and μ. To make the learning feasible within experimental constraints and to maintain interpretability, we use a simple tile coding scheme to discretize our state space: continuous values of a_z and ω are each discretized into three states (+, 0, −), partitioned by thresholds $\pm K_a$ and $\pm K_\omega$ respectively. The thresholds are set at ± 0.8 times the standard deviation of a_z and ω. Because the width of the distributions of a_z and ω can vary across days, the data obtained on a particular day are normalized by the standard deviation calculated for that day. In effect, the filtration threshold to detect a signal against turbulent 'noise' is higher on days with more turbulence. The consequence is that the behaviour of the learned strategy could change across days, adapting to the recent statistics of the environment. The bank angle takes five possible values $(0°, \pm 15°, \pm 30°)$, while the three possible actions allow for increasing, decreasing by 15° or keeping the same bank angle. In summary, we have a total of $3 \times 3 \times 5 = 45$ states in the state space and three actions in the action space.

We choose the local vertical wind acceleration a_z obtained in the next time step as the reward function. The choice of a_z as an appropriate reward signal is motivated by observations made in simulations from ref. [20]. In Supplementary Information, we show that the obtained policy using a_z as the reward function is equivalent to a policy that also maximizes the expected gain in height.

Learning the strategy in the field. Data collected in the field are split into (s, a, s', r) quadruplets containing the current state s, the current action a, the next state s' and the obtained reward r, which are pooled to obtain the transition matrix $T(s' \mid s, a)$ and reward function $R(s, a)$. Value iteration methods are used to estimate the Q values from T and R. The learning process is offline and off-policy; specifically, we begin training with a 'random' policy that takes the three possible actions with equal probability irrespective of the current state. This behavioural policy was used for 12 out of the 15 days of training. For the other days, a softmax policy[7] with 'temperature' parameter set to 0.3 was used. For softmax training, the Q values were first estimated from the data obtained in the previous days and then normalized by the difference between the maximum and minimum Q values over the three possible actions at a particular state, as described in ref. [20].

Using a fixed, random policy as our behavioural policy has the disadvantage that it slows learning, as state–action pairs that rarely appear in the final policy are still sampled. On the other hand, it has the benefit that calibrating the parameters necessary for the unbiased measurement of a_z and ω (see Supplementary Information) is performed simultaneously with learning, which considerably reduces the number of days required in the field. Importantly, offline learning permits us to continuously monitor the variance of the estimated Q values by bootstrapping from the set E of accumulated (s, a, s', r) quadruplets up to a particular point. Specifically, $|E|$ samples are drawn with replacement from E, and Q values are obtained for each state–action pair by value iteration. The steps are repeated and the average of the bootstrapped standard deviations in Q over all the state–action pairs is used as a measure of learning progress, as shown in Fig. 2a.

We expect certain symmetries in the transition matrix and the reward function, which we exploit to expedite our learning process. Particularly, we note that the Markov decision process is invariant to an inversion of sign in the bank angle $\mu \to -\mu$. This transforms a state as $(a_z, \omega, \mu) \to (a_z, -\omega, -\mu)$ and inverts the action

from that of increasing the bank angle to decreasing the bank angle and vice versa. We symmetrize T and R as

$$T^{\mathrm{sym}} = \frac{T^{+} + T^{-}}{2} \tag{9}$$

$$R^{\mathrm{sym}} = \frac{R^{+} + R^{-}}{2} \tag{10}$$

where $+$ and $-$ denote the obtained values and those computed by applying the inverting transformation respectively. Finally, T^{sym} and R^{sym} are used to obtain a symmetrized Q function, which results in a symmetric policy as shown in Fig. 2b. To conveniently obtain the policy that uses only a_z (Fig. 3d), the above procedure is repeated with the threshold for ω (K_ω) set to infinity.

Testing the performance of the learned policy in the field. To obtain the data shown in Fig. 3b, the glider is first sent autonomously to an arbitrary but fixed location 250 m above ground level. The learned policy for thermals is then turned on, and the mean climb rate (that is, the total height gained divided by the total time) is measured over a 3-min interval. To obtain the control data, the glider instead follows a random policy, which takes the three possible actions with equal probability. Trials in which we observe little to no atmospheric convection are filtered out by imposing a threshold on the standard deviation of the vertical wind velocity over the 3-min trial. In Extended Data Fig. 6, we show the distribution of the standard deviation in w_z collected from about 240 3-min trials over 9 days. Trials below the threshold chosen as the 25th percentile mark (red dashed line) are not used for our analysis.

Testing performance for different wingspans in simulations. Soaring performance is analysed in simulations similar to those developed in ref. [20] and adapted to reflect the constraints faced by our glider and the environments typically observed in the field.

The atmospheric model consists of two components: (1) a kinematic model of turbulence that reproduces the statistics of wind velocity fluctuations in the convective atmospheric boundary layer; and (2) the positions, sizes and strengths of updrafts and downdrafts. The temporal and spatial statistics of the generated velocity field satisfy the Kolmogorov and Richardson laws[29] and the mean velocity profile in the convective boundary layer[5], as described in the supplementary information of ref. [20]. Stationary updrafts and downdrafts of Gaussian shape are placed on a staggered lattice of spacing approximately 125 m on top of the fluctuating velocity field. Specifically, their contribution to the vertical wind velocity at position r is given by

$$w_z = \pm W \mathrm{e}^{-(r_\perp - r_\perp^0)^2/(2R^2)} \tag{11}$$

where $r_\perp^0$ is the location of the centre of the up(down)draft in the horizontal plane, W is its strength and R is its radius. W is drawn from a half-normal distribution of scale 1.5 m s^{-1}, whereas the radius is drawn from a (positive) normal distribution of mean 40 m and deviation 10 m. Gaussian white noise of magnitude 0.2 m s^{-1} is added as additional measurement noise.

We assume that the glider is in mechanical equilibrium; the lift, drag and weight forces on the glider are balanced, except for centripetal forces while turning. The parameters corresponding to the lift and drag curves and the (fixed) angle of attack are set such that the airspeed is $V = 8$ m s^{-1} and the sink rate is 0.9 m s^{-1} at zero bank angle, which match those measured for our glider in the field. Control over bank angle is similar to those imposed in the experiments: that is, the bank angle switches linearly between the angles 0°, ±15°, ±30° in a time interval t_a, corresponding to the time step between actions. The glider's trajectory and wind velocity readings are updated every 0.1 s. The vertical wind acceleration is derived assuming that the glider directly reads the local vertical wind velocity. The gradients in vertical wind velocity across the wings are estimated as the difference between the vertical wind velocities at the two ends of the wings. The readings are smoothed with exponential smoothing kernels; the smoothing parameters in experiments are chosen to coincide with those that yield the greatest height gain in simulations.

Estimation of noise in gradient sensing due to atmospheric turbulence. The cues a_z and ω measure the gradients in the vertical wind velocity along and perpendicular to the heading of the glider. Updrafts and downdrafts are relatively stable structures in a varying turbulent environment. Thermal detection through gradient sensing constitutes a discrimination problem of deciding whether a thermal is present or absent given the current a_z and ω. We estimate the magnitude of turbulent 'noise' that unavoidably accompanies gradient sensing. Intuitively, turbulent fluctuations in the atmospheric boundary layer (ABL) are made up of eddies of different length scales, with the largest being the size of the height of the ABL. Energy is transferred from larger, stronger eddies to smaller, weaker eddies, and eventually dissipates at the centimetre scale owing to viscosity in the bulk and owing to the boundary at the Earth's surface. In Supplementary Information, we present an explicit calculation of the signal-to-noise ratio for ω estimation, taking into account the effect of turbulent eddies on the statistics of noise. Below, we give simple scaling arguments and refer to Supplementary Information for further details.

A glider moving at an airspeed V and integrating over a timescale T averages a_z over a length VT. For V much larger than the velocity scale of the eddies, which is typically the case, the decorrelation of wind velocities is due to the glider's motion; the eddies themselves can be considered to be frozen in time. The magnitude of the spatial fluctuations across the eddy of this size scales according to the Richardson–Kolmogorov law[29] as $(VT)^{1/3}$. The mean gradient signal when going up the gradient scales as (VT); the resultant signal-to-noise ratio in a_z scales as $(VT)^{2/3}$.

Similar arguments are applicable for ω measurements. In this case, the signal-to-noise ratio has an additional dependence on the wingspan l. The dominant contribution to the noise comes from eddies of size l, whose strength scales as $l^{1/3}$. As the glider moves a distance VT, for $l \ll VT$, it traverses VT/l distinct eddies of size l. Consequently, the noise is averaged out by a factor $(VT/l)^{-1/2}$, corresponding to the VT/l independent measurements. Multiplying these two factors, the averaged noise is proportional to $l^{5/6}(VT)^{-1/2}$. As the mean gradient (that is, the signal) is approximately l, the signal-to-noise ratio is then proportional to $l^{1/6}(VT)^{1/2}$.

From the above arguments and dimensional considerations, we get order-of-magnitude estimates of the signal-to-noise ratio (SNR) for a_z and ω estimation:

$$\mathrm{SNR}_{a_z} \propto \frac{WV^{2/3}T^{2/3}L^{1/3}}{wR} \tag{12}$$

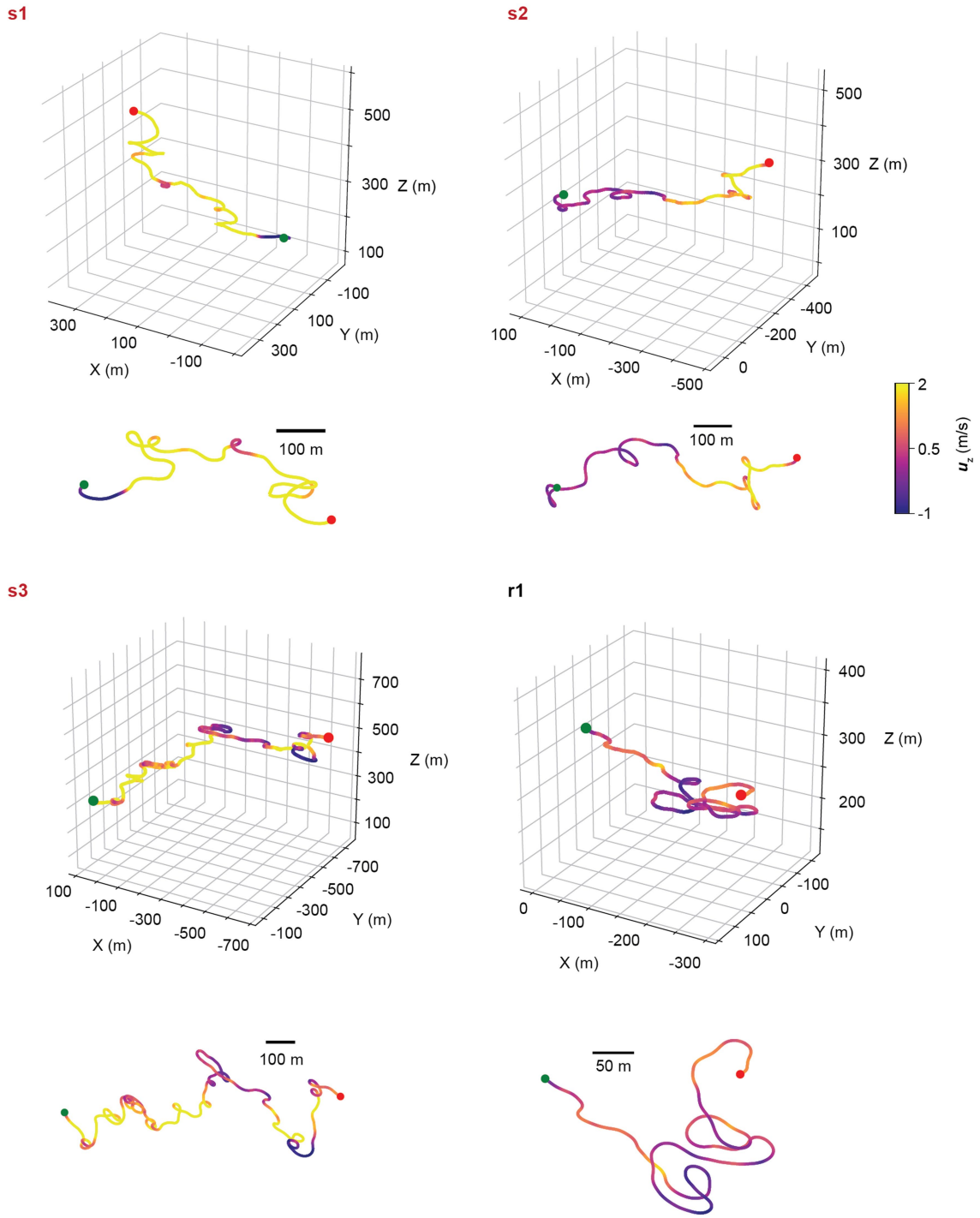
$$\mathrm{SNR}_{\omega} \propto \frac{WV^{1/2}T^{1/2}l^{1/6}L^{1/3}}{wR} \tag{13}$$

where W is the strength of the thermal, R is its radius, w is the magnitude of turbulent vertical wind velocity fluctuations and L is the length scale of the ABL. For the signal-to-noise ratio estimates presented in the text, we use $W = 2$ m s^{-1}, $R = 50$ m, $l = 2$ m, $V = 8$ m s^{-1}, $T = 3$ s, $L = 1$ km. The values of V and T correspond to the airspeed of the glider in experiments and the timescale between actions during learning respectively.

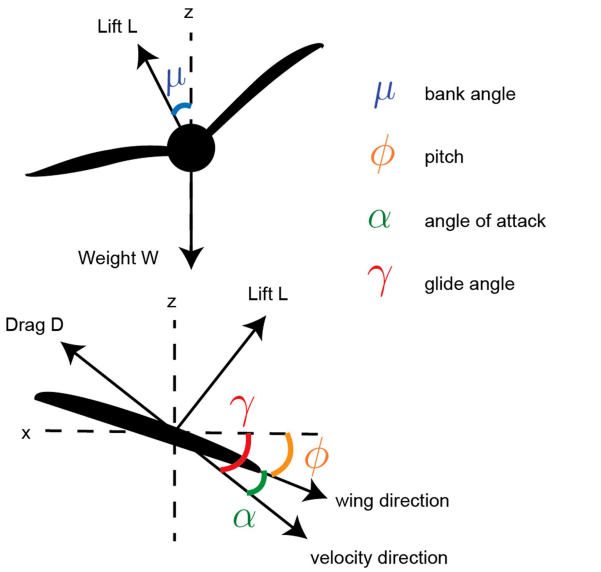
Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

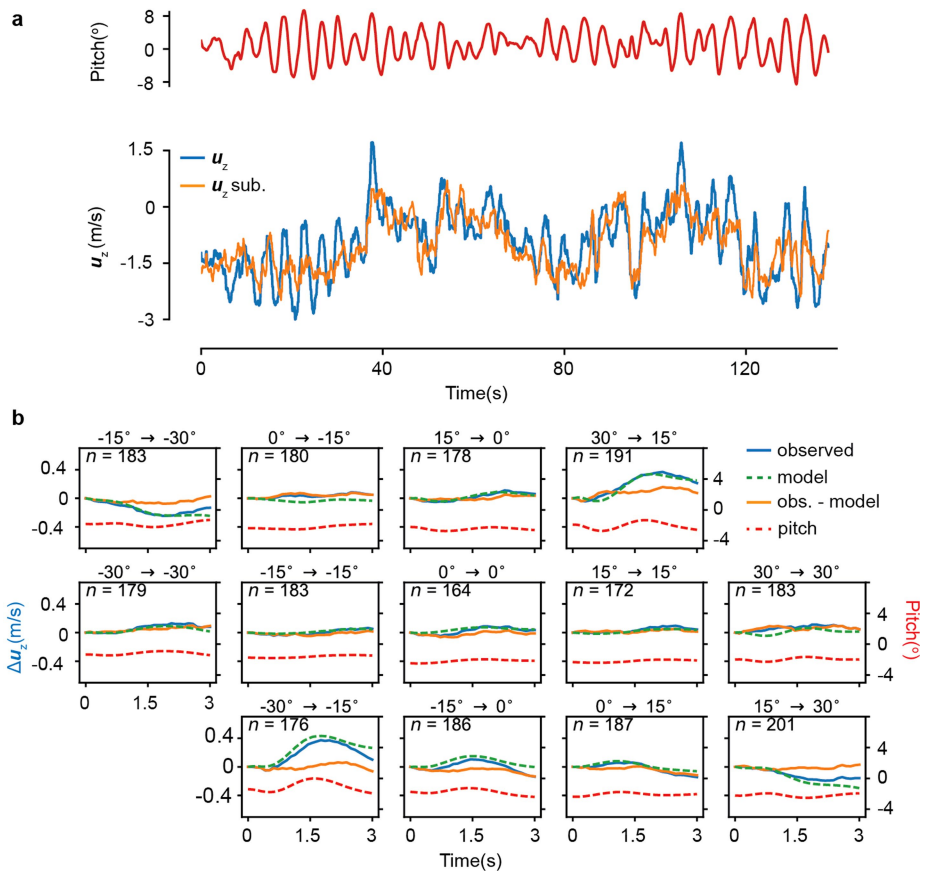
29. Frisch, U. *Turbulence: The Legacy of A. N. Kolmogorov* (Cambridge Univ. Press, Cambridge, 1995).



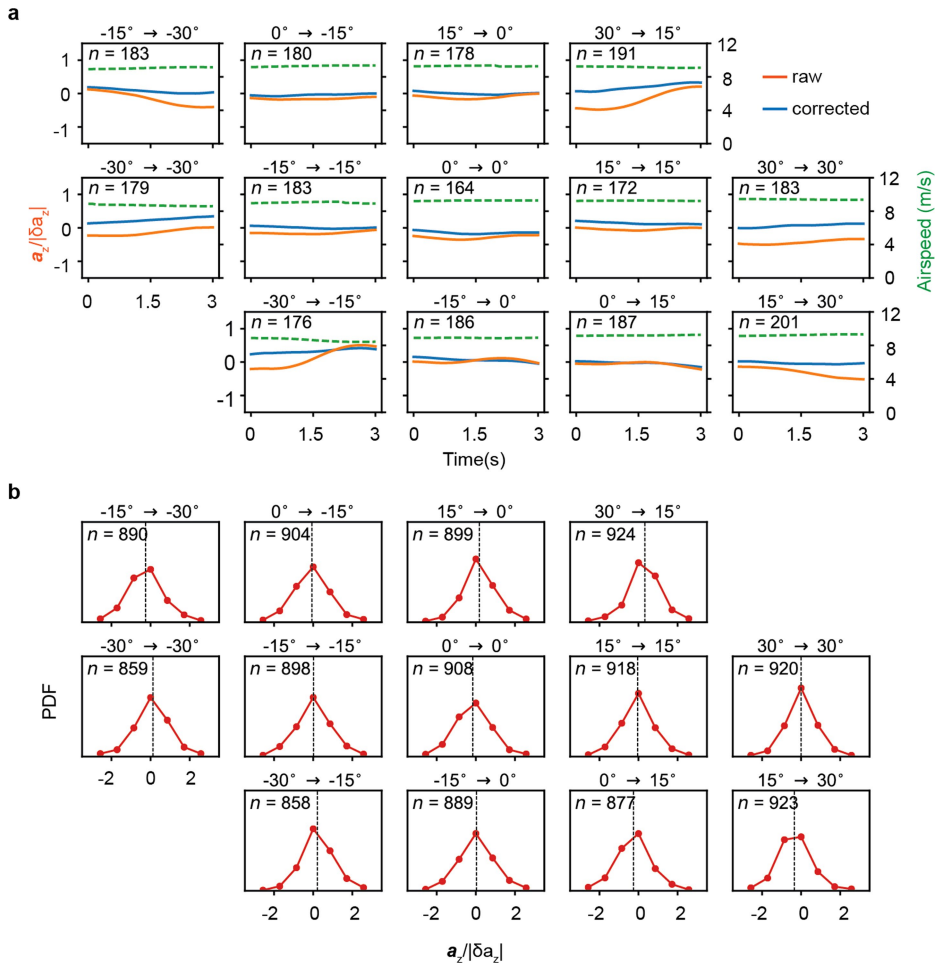
Extended Data Fig. 1 | Sample trajectories obtained in the field. The three-dimensional view and top view are shown of the glider's trajectory as it executes the learned strategy for thermals (labelled 's') or a random policy that takes actions with equal probability (labelled 'r'). The trajectories are coloured according to the instantaneous vertical ground velocity u_z. The green (red) dot shows the start (end) point of the trajectory. Trajectories s1, s2 and r1 last for 3 min each, whereas s3 lasts for about 8 min.



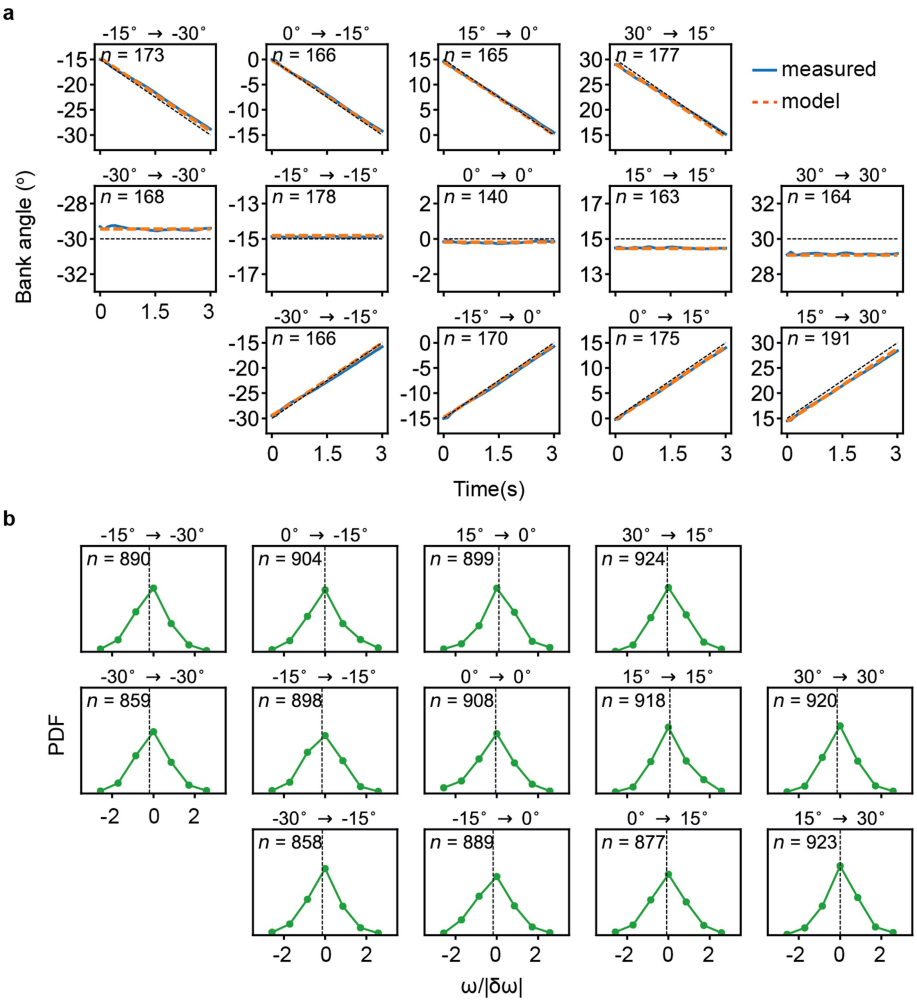
Extended Data Fig. 2 | Force–body diagram of a glider. The forces on a glider and the definitions of the various angles that determine the glider's motion.



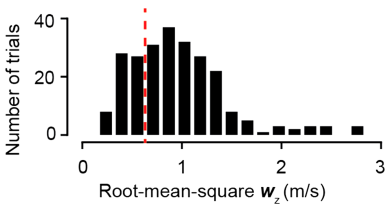
Extended Data Fig. 3 | Modelling the longitudinal motion of the glider. **a**, Sample trajectory of a glider's pitch and its vertical velocity with respect to ground (u_z) in a case in which the feedback control over the pitch is reduced in order to exaggerate the pitch oscillations. The blue line shows the measured u_z, and the orange line is u_z obtained after subtracting the contributions from longitudinal motions of the glider (see Supplementary Information). **b**, The blue line shows the average change in u_z when a particular action is taken (labelled above each panel), averaged over n 3-s intervals. The 13 panels correspond to the 13 possible bank angle changes from the angles 0°, ±15° and ±30° by increasing, decreasing the bank angle by 15° or keeping the same angle. The green dashed line shows the prediction from the model whereas the orange line is the estimated w_z. The axis on the right shows the averaged pitch (red dashed line).



Extended Data Fig. 4 | The estimated vertical wind acceleration is unbiased after accounting for the glider's longitudinal motion. **a**, The averaged vertical wind acceleration a_z in units of its standard deviation. a_z, plotted as in Extended Data Fig. 3b, is shown in orange with (blue line) and without (orange line) accounting for the glider's longitudinal motions. The axis on the right shows the airspeed (green dashed line). **b**, Probability density functions (PDFs) of a_z for the different bank angle changes. The black dashed line shows the median.



Extended Data Fig. 5 | The estimated roll-wise torque is unbiased after accounting for the effects of feedback control and glider aerodynamics. **a**, The averaged evolution of the bank angle shown as in Extended Data Fig. 3b. The blue line shows the measured bank angle and the dashed orange line shows the best-fit line obtained from simultaneously fitting the 13 blue curves to the prediction (see Supplementary Information). **b**, PDFs of the roll-wise torque ω (in units of its standard deviation) for the different bank angle changes. The black dashed line shows the median value.



Extended Data Fig. 6 | The distribution of the strength of vertical currents observed in the field. The root-mean-square vertical wind velocity measured in the field is pooled from about 240 3-min trials collected over 9 days. The dashed red line shows the threshold criterion imposed when measuring the performance of the strategy in the field (see Methods).

Extended Data Table 1 | Parameter values

Label	Description	Value
l	Wingspan of glider used in experiments	2m
φ_d	Desired pitch	-2°
τ	Feedback control time scale	0.45s
t_a	Interval between actions (learning)	3s
t_a	Interval between actions (soaring)	1.5s
$\alpha_0 - \alpha_i$	Net angle of attack (see eq. 6)	14°
V	Airspeed (typical)	6 to 8 m/s
T_{dih}	Dihedral effect timescale (typical)	14 to 30 s
T_{ob}	Overbanking effect timescale (typical)	< -20s
b	Trim bias (typical)	-2 to +2°/s
T_{roll}	Opposing roll timescale (typical)	1.5 to 3 s
$\pm K_a$, $\pm K_w$	Thresholds for $\boldsymbol{a}_z$ and ω state estimation	0.8 x std. dev
σ_a, σ_a'	Exponential smoothing timescales for $\boldsymbol{a}_z$	$8t_a/3$, $2t_a/3$
σ_w, σ_w'	Exponential smoothing timescales for ω	t_a, $t_a/4$
γ	Discount factor for RL implementation	0.8