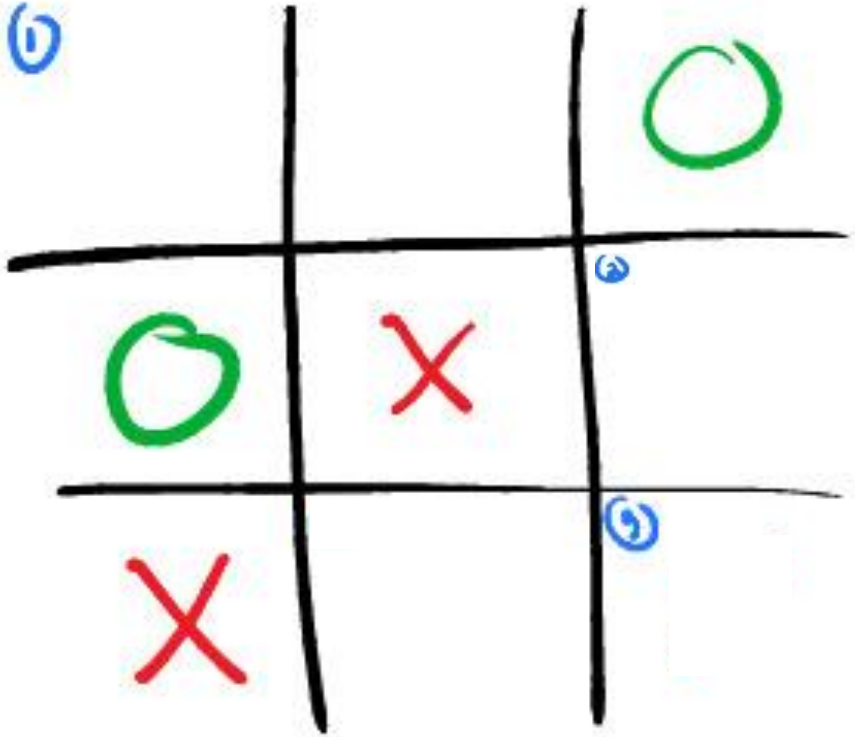
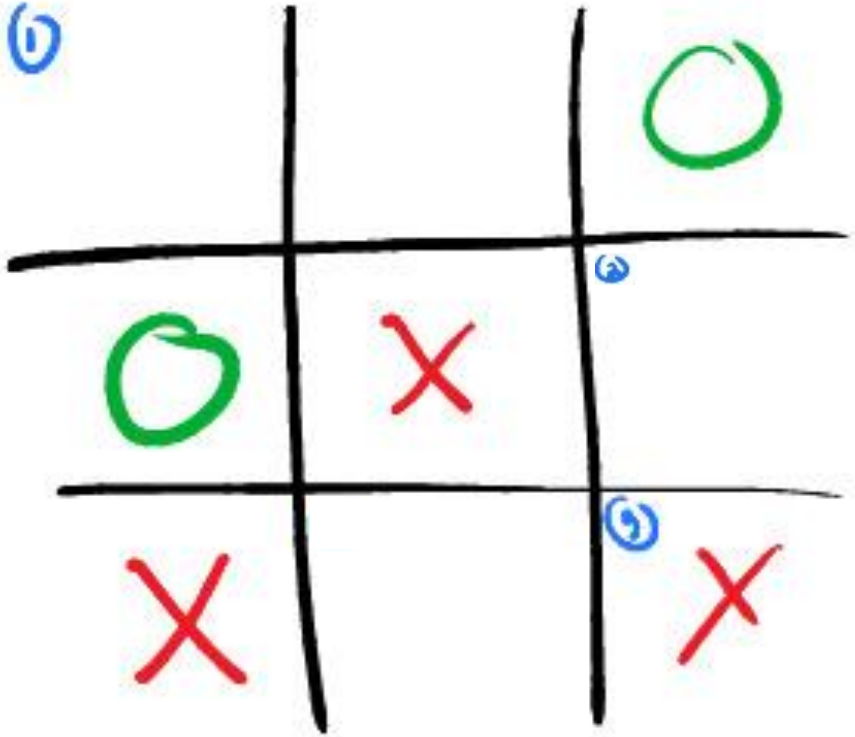
Decisions from data
Controlling complex systems with reinforcement learning

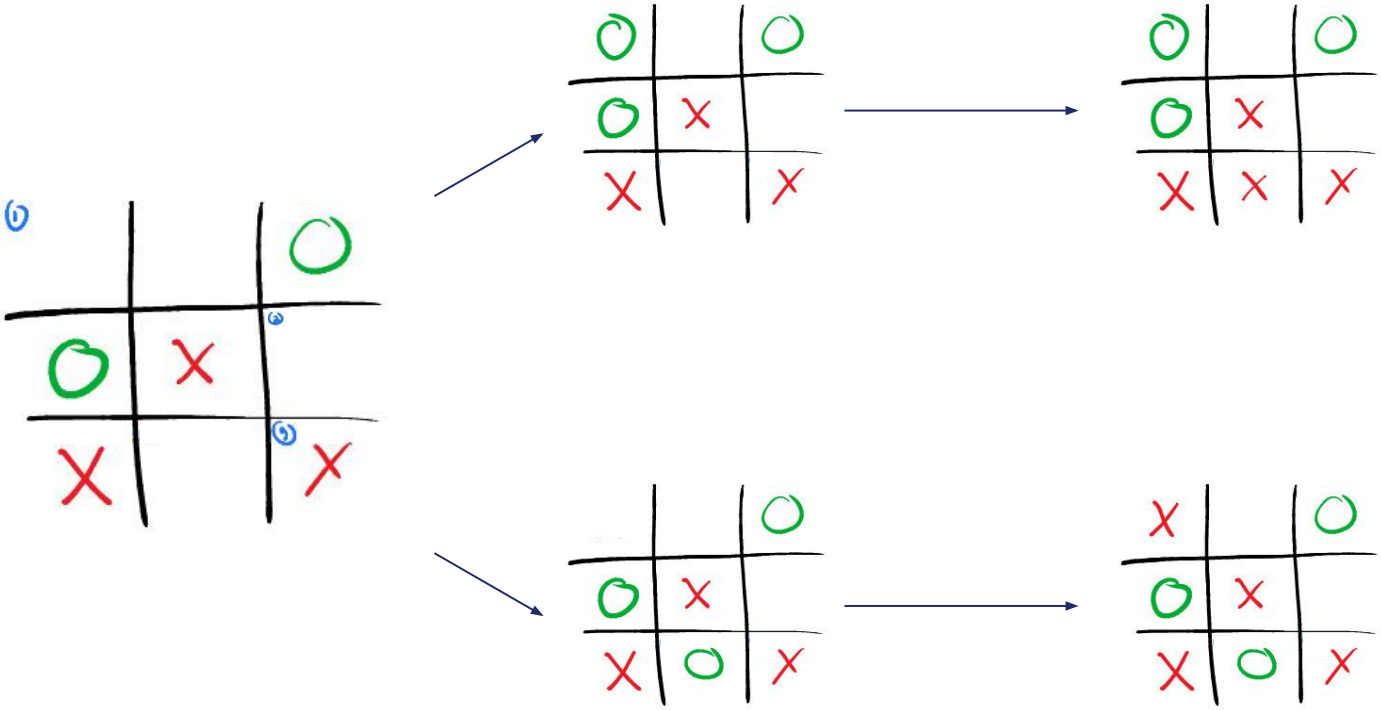
Marc G. Bellemare
Google Brain, Montreal; Mila

With: Salvatore Candido, Pablo Samuel Castro, Jun Gong, Marlos C. Machado, Subhodeep Moitra, Sameera S. Ponda, Ziyu Wang

IVADO







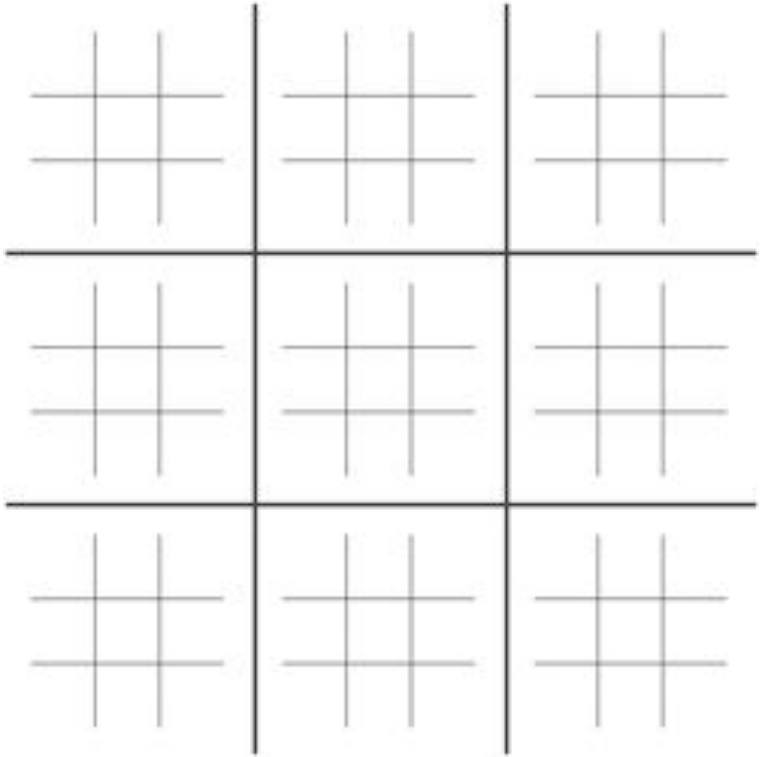


Image credit: https://bejofo.net/ttt

Reinforcement learning = trial and error

data → decisions

Reinforcement learning = trial and error

data → decisions

Trial and error and cats

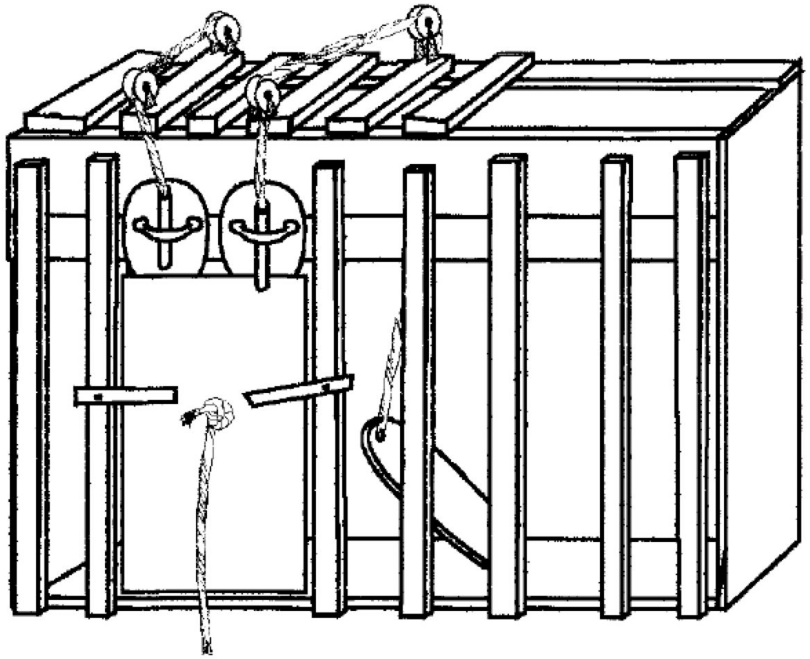


Fig. 4. Box K. The door is held in place by a weight suspended by a string. To open the door, a cat had to depress a treadle, pull on a string, and push a bar up or down. (After Thorndike, 1898, Figure 1, p. 8.)

Credit assignment is at the heart of RL

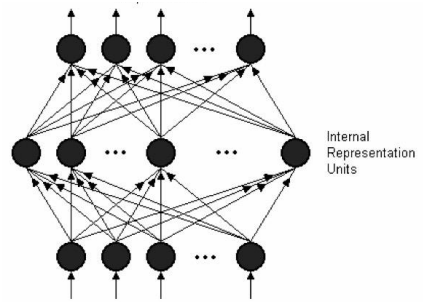
Credit assignment via the Bellman equation

Markov decision process

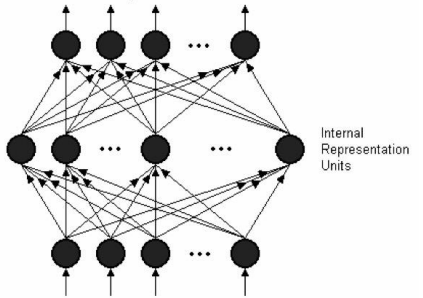
$$Q^*(x,a) = R(x,a) + \gamma \mathop{\mathbf{E}}_{x' \sim P} \left[\max_{a' \in \mathcal{A}} Q^*(x',a') \right]$$

Implemented as a
Deep neural network

Deep reinforcement learning



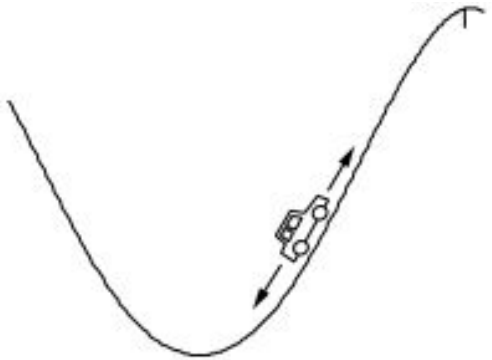
80 units; Tesauro (1995)



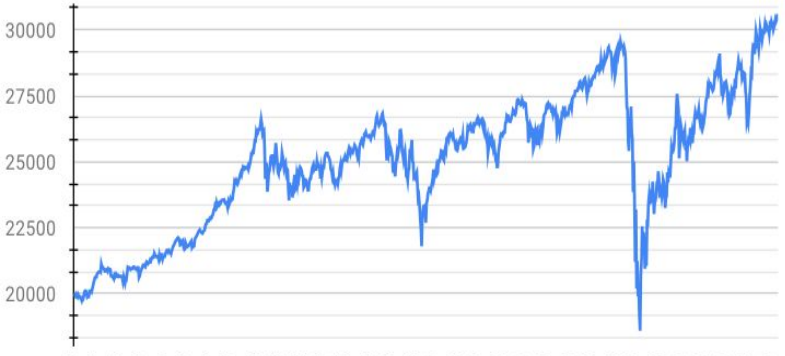
40 x 256 convolutional filters;
Silver et al. (2017)

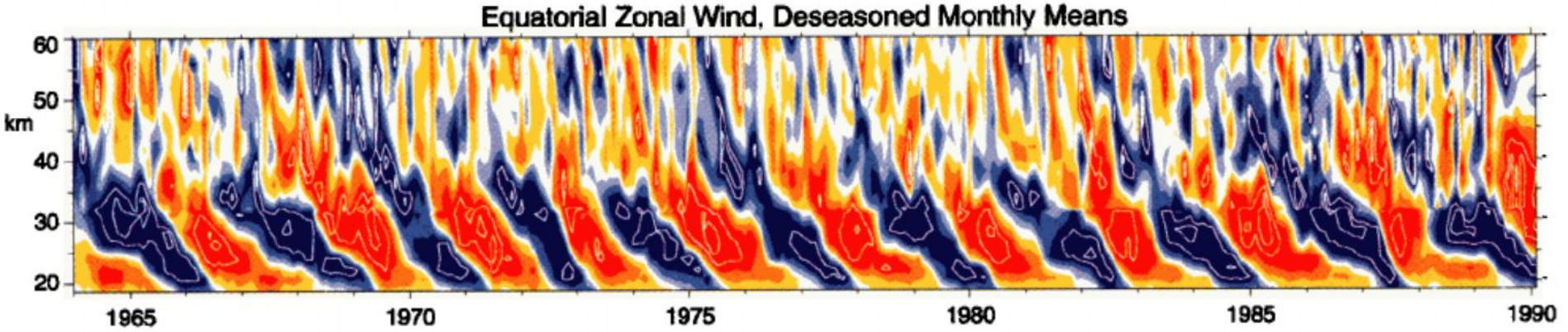
Many RL problems are...

Underactuated

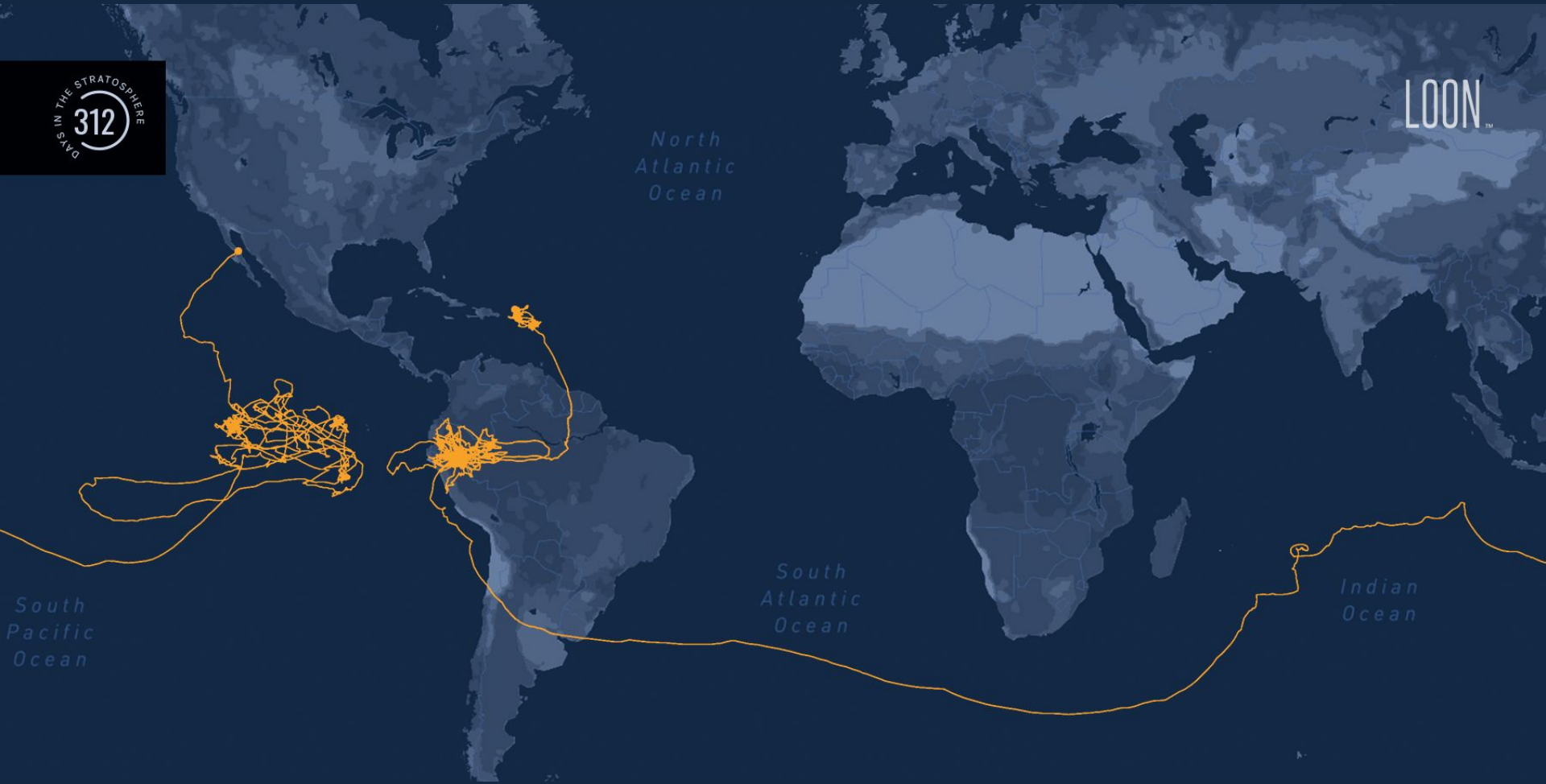


Partially observable

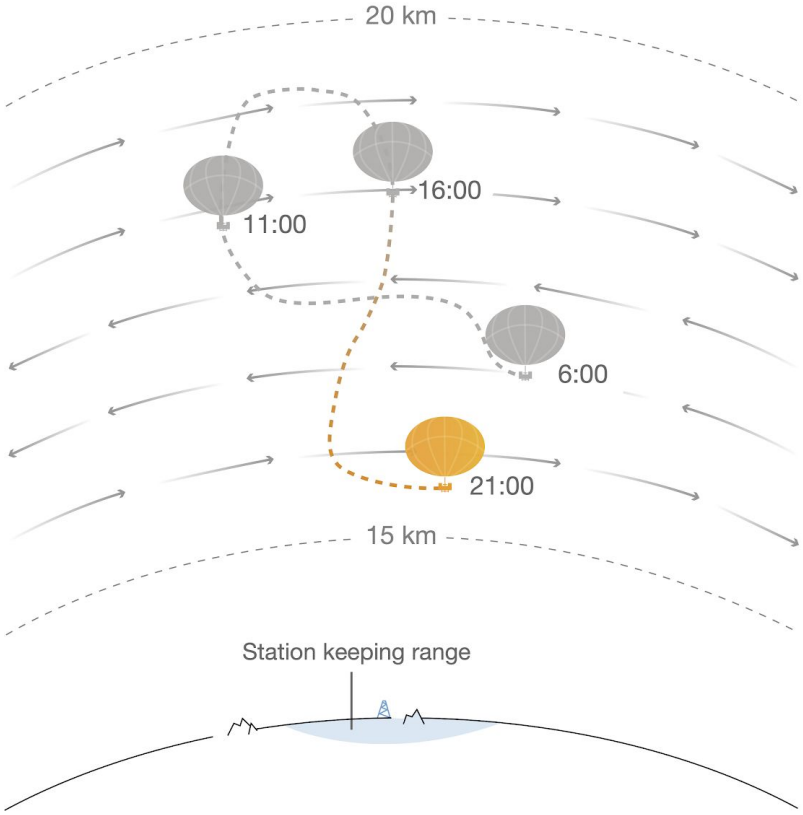


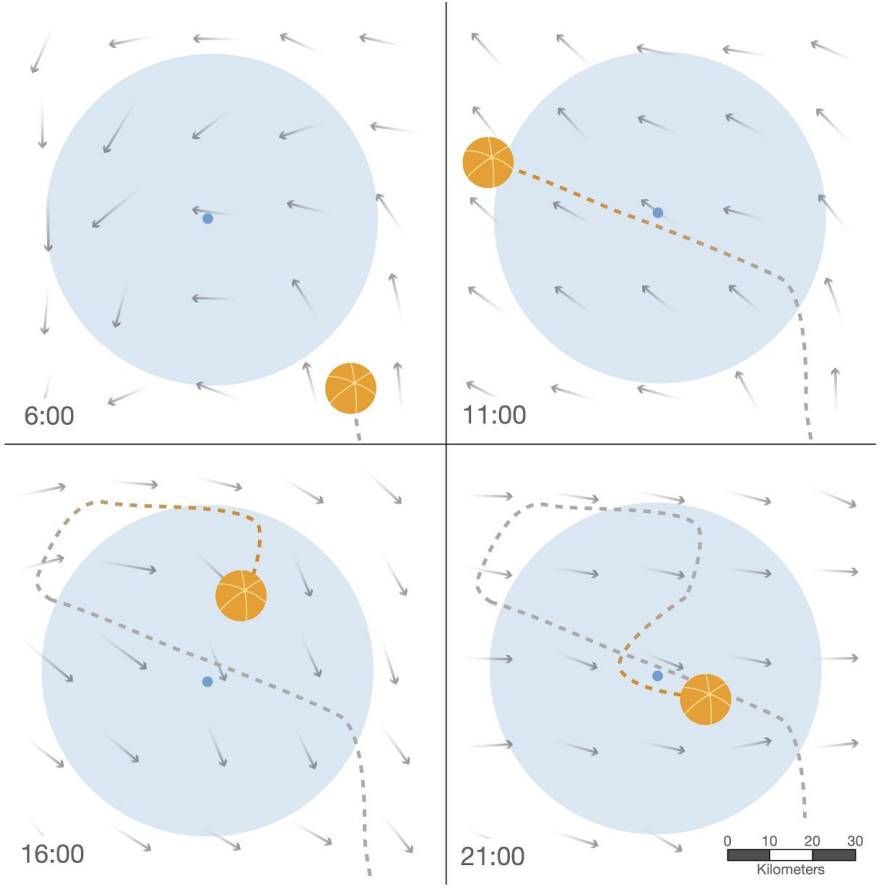


The quasi-biennial oscillation, Baldwin et al. (2001)



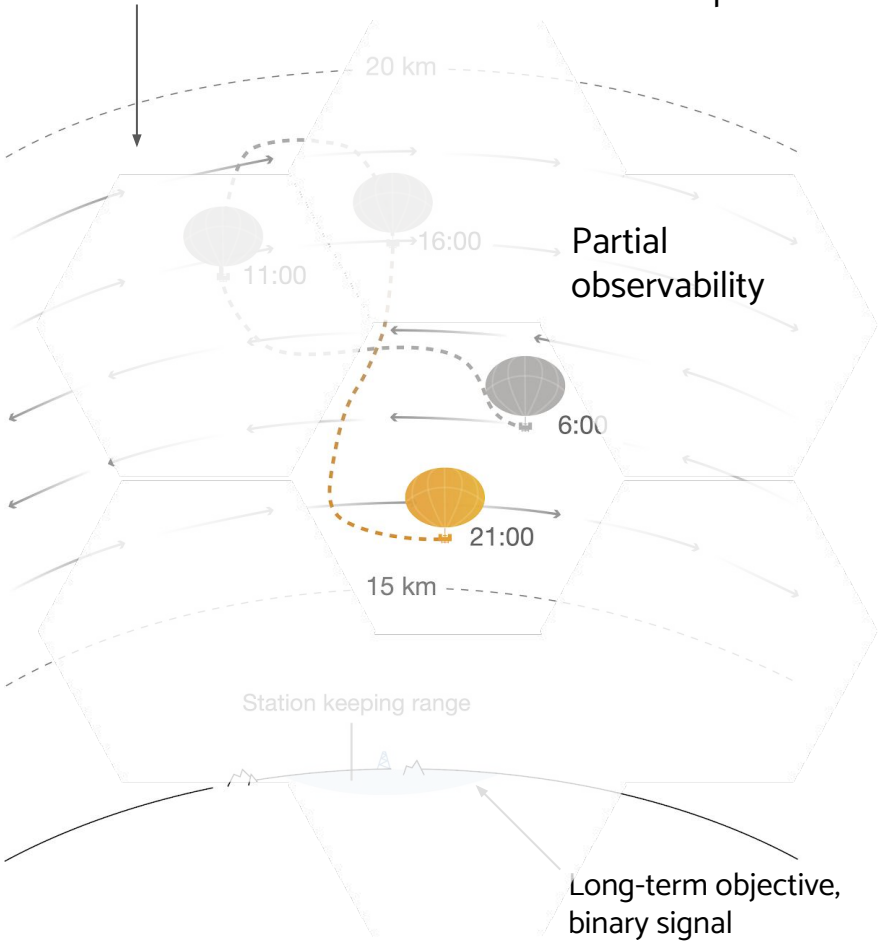
312 Days in the Stratosphere, Loon, Oct 28 2020.


20 km
11:00
16:00
6:00
21:00
15 km
Station keeping range


6:00
11:00
16:00
21:00
0 10 20 30
Kilometers

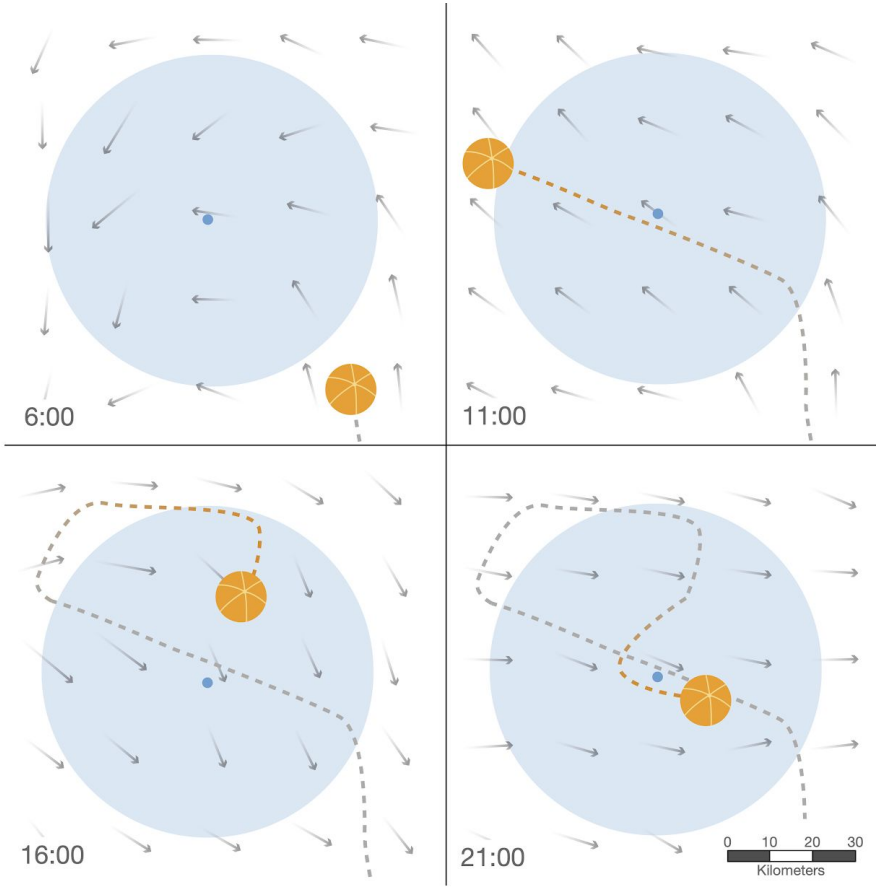
Underactuated system,
stochastic dynamics

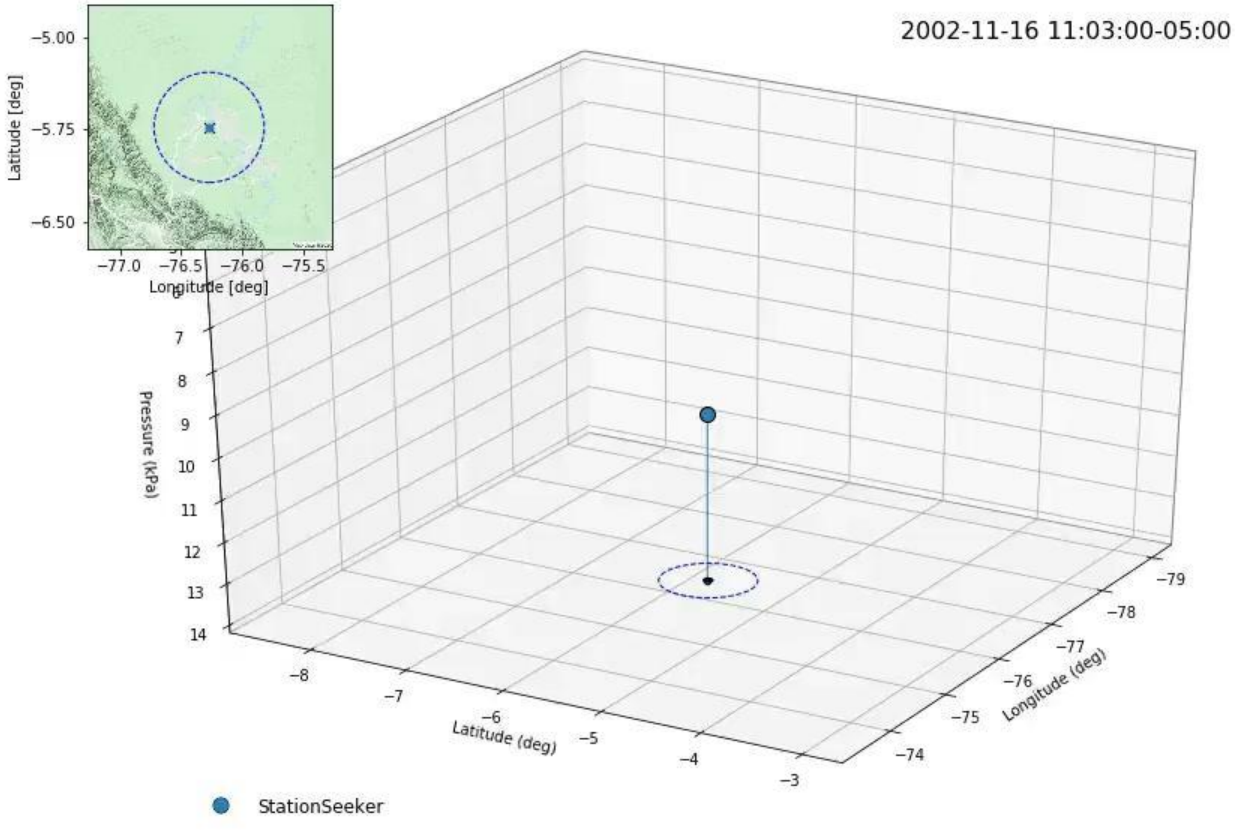
Limited
power



Partial
observability

Long-term objective,
binary signal




2002-11-16 11:03:00-05:00
Latitude [deg]
−5.00
−5.75
−6.50
−77.0
−76.5
−76.0
−75.5
Longitude [deg]
Pressure (kPa)
7
8
9
10
11
12
13
14
−8
−7
−6
−5
−4
−3
Latitude (deg)
−79
−78
−77
−76
−75
−74
Longitude (deg)
StationSeeker

StationSeeker in equations

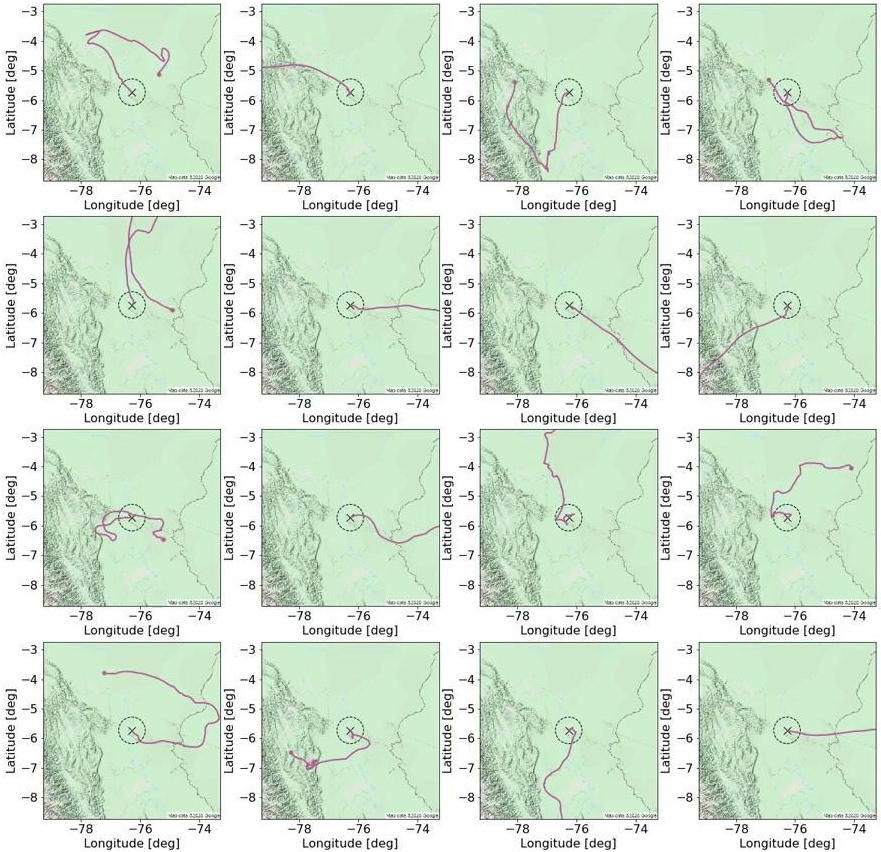
1) Wind score.

$$g_l = (1 - \alpha_\Delta)e^{-w_\Delta \theta_l} + \alpha_\Delta e^{-k_1 \mu_l}$$

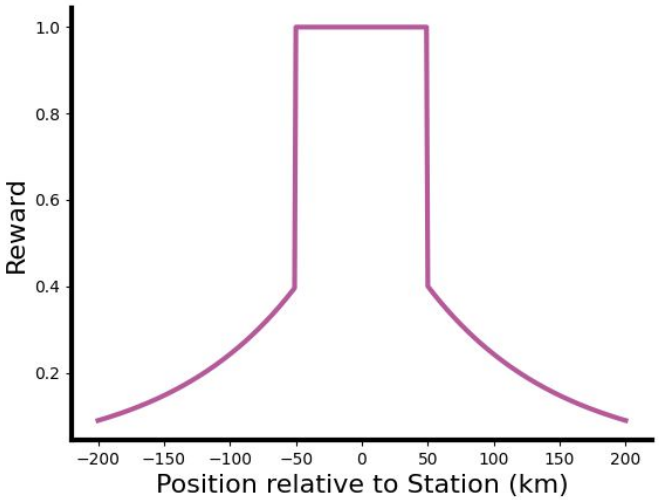
2) Per-altitude score.

$$s_l = (1 - u_l)g_l + u_l g_{\text{unknown}} + k_2 e^{-k_3 |l - l_{\text{current}}|}$$

3) Setpoint to max. scoring altitude.

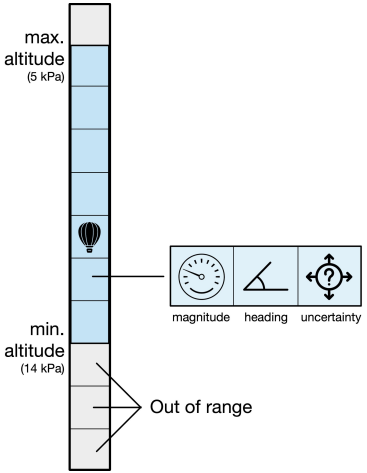

Latitude [deg]
−3
−4
−5
−6
−7
−8
Longitude [deg]
−78
−76
−74

Deep reinforcement learning for balloon navigation



$$r(\Delta) = c_{\text{CLIFF}} 2^{-(\Delta-\rho)/\tau} \times \text{POW}(\omega)$$

Forecast + measurements + Gaussian process = **wind column**

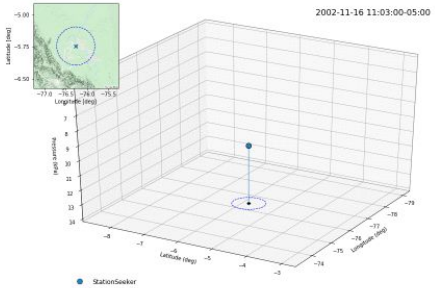


+16 ambient variables

The simulator

The ERA5 reanalysis (dataset) provides **baseline winds**

- Like real, but
- Low resolution.

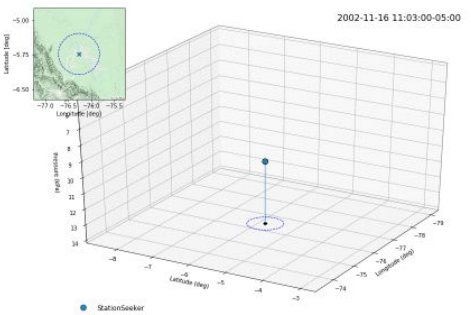


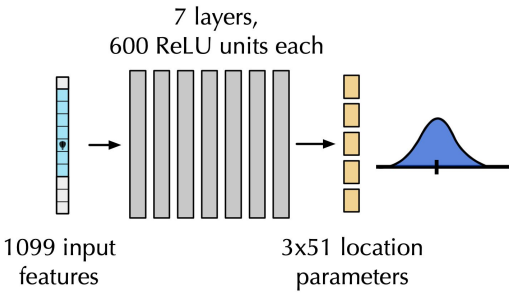
Baseline winds are **upsampled** using procedural noise:

- Statistically plausible
- High resolution
- Effectively infinite supply

An **episode** is defined by **initial conditions** *(latitude, longitude, altitude, time),* station location, and random seed

Design and training





2-day training simulations

In the tropics (+/- 25 lat.)

Starting up to 200km away

Light filtering of “impossible” conditions

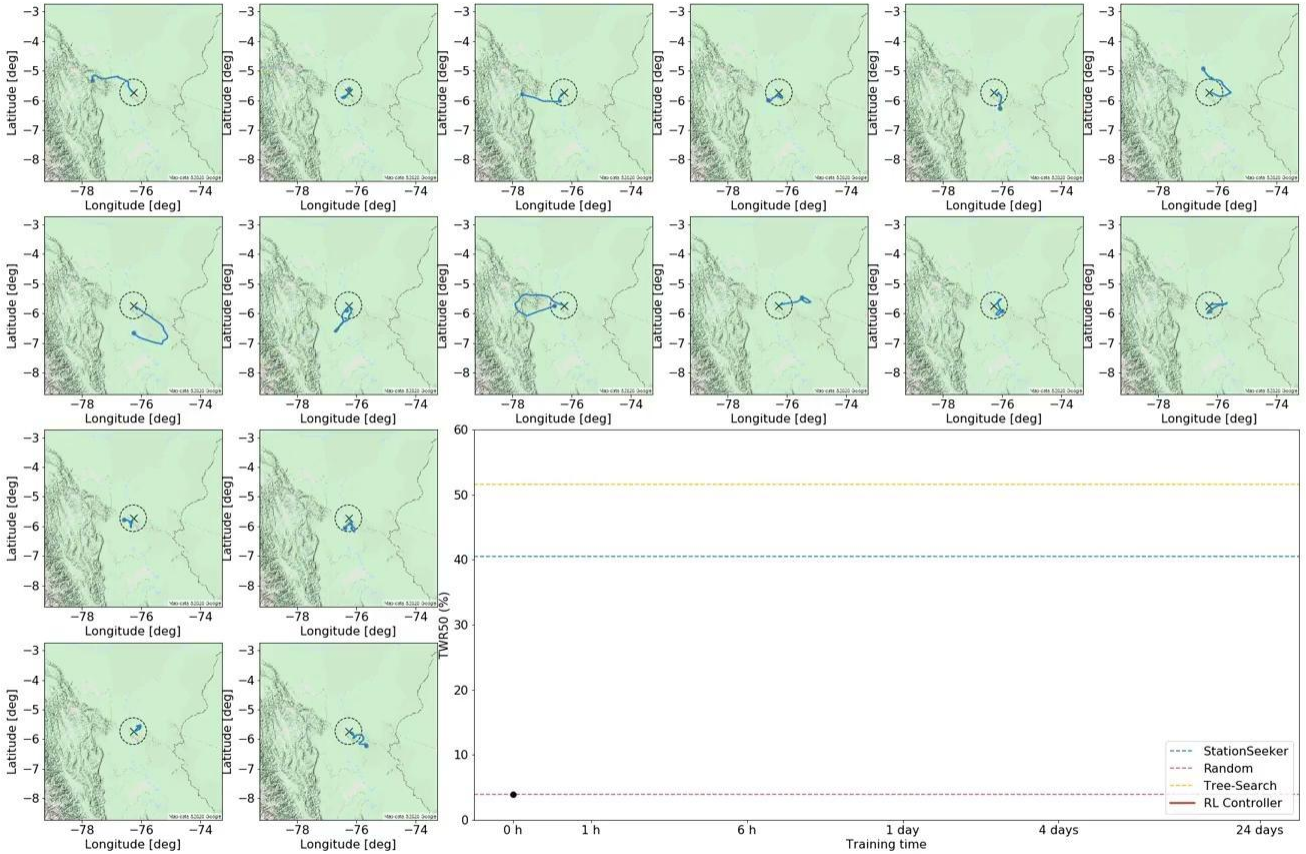
Combined exploration + greedy behaviour

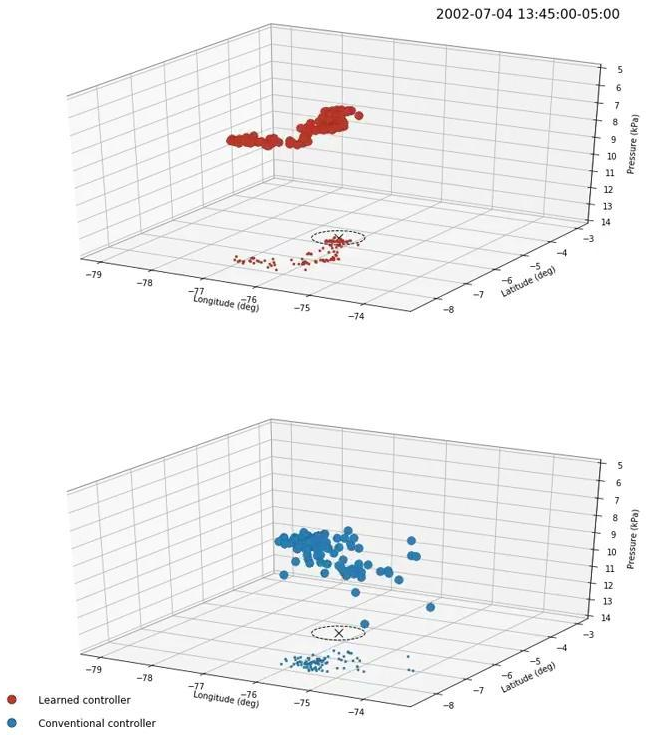
Distributional predictions (QR-DQN)

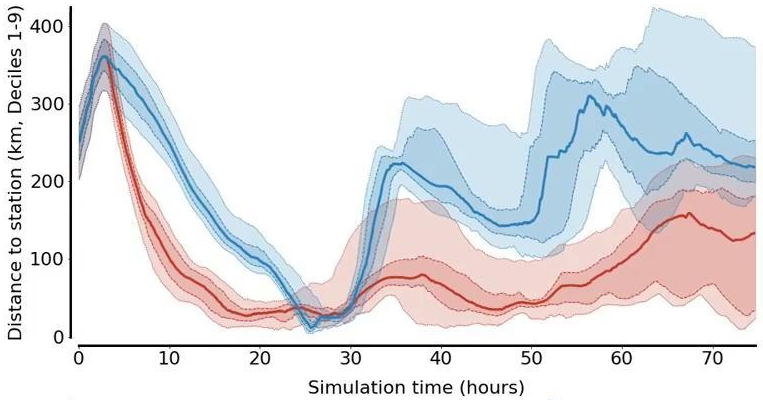
Distributed training:

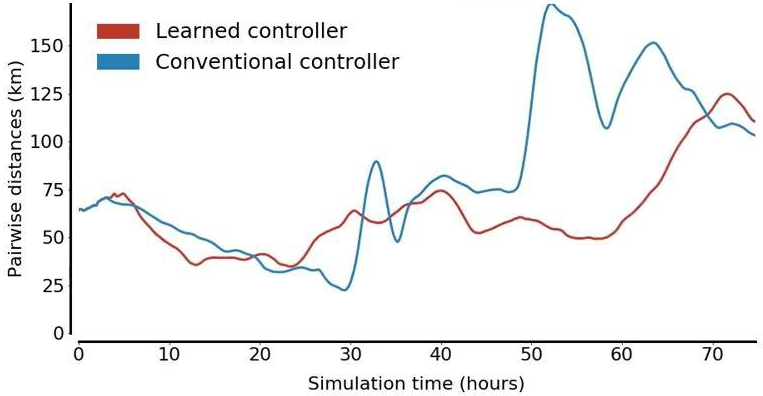
- 100 actors
- 4 replay buffers
- 1 GPU

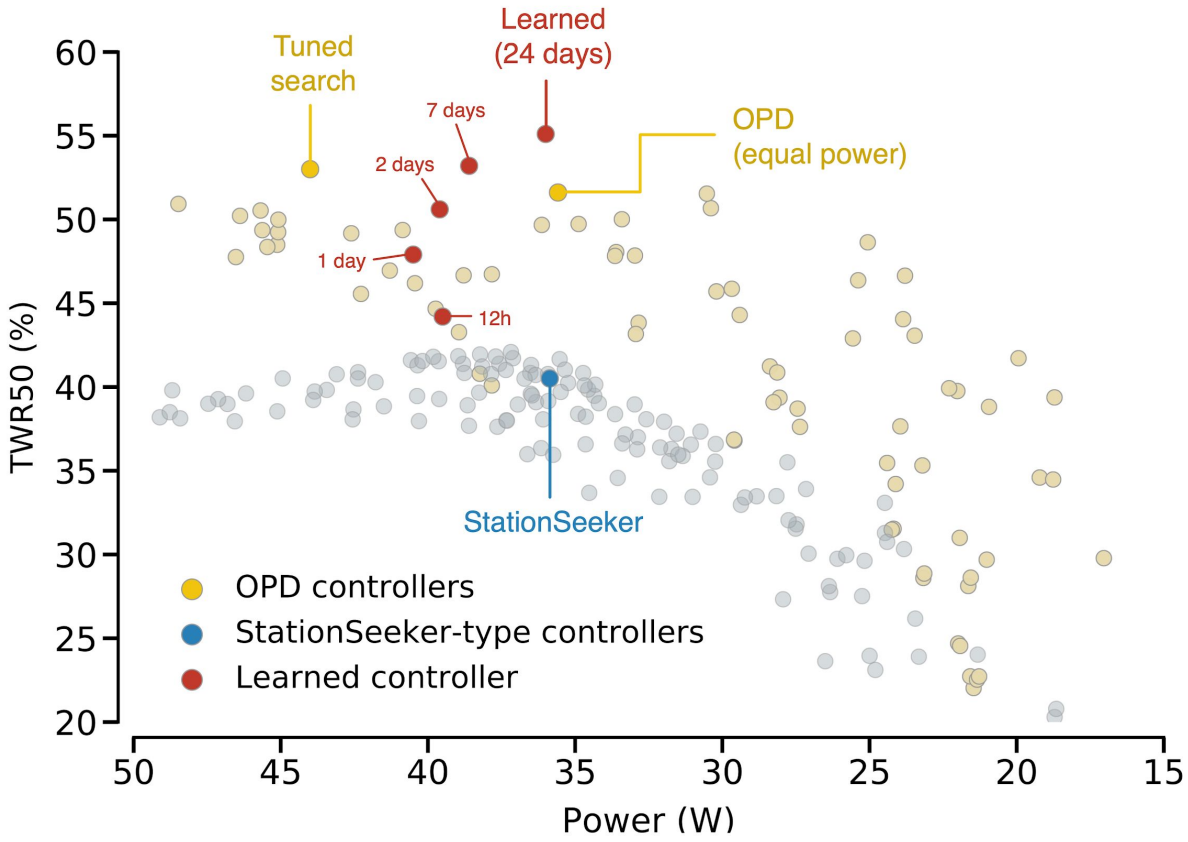
1.1B training steps (~30 days wall time)

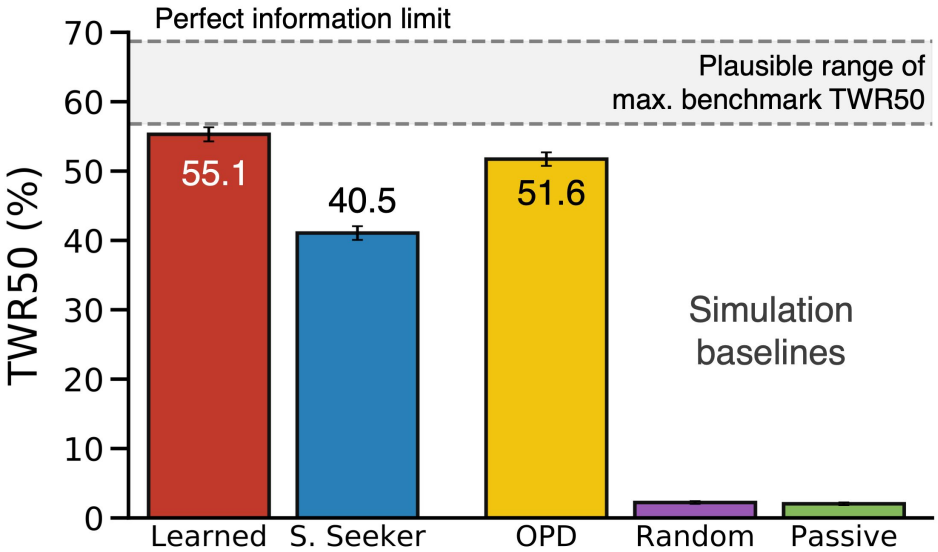

Latitude [deg]
Longitude [deg]
TWR50 (%)
Training time
0 h
1 h
6 h
1 day
4 days
24 days
StationSeeker
Random
Tree-Search
RL Controller

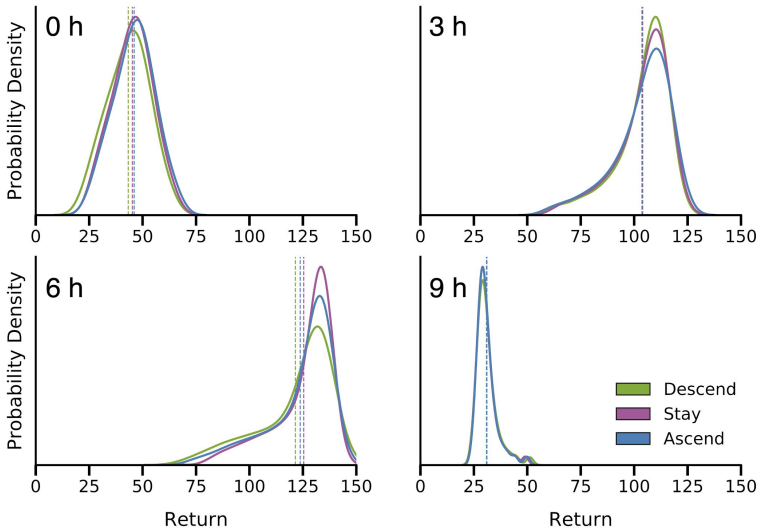

2002-07-04 13:45:00-05:00
Pressure (kPa)
Longitude (deg)
Latitude (deg)
Learned controller
Conventional controller

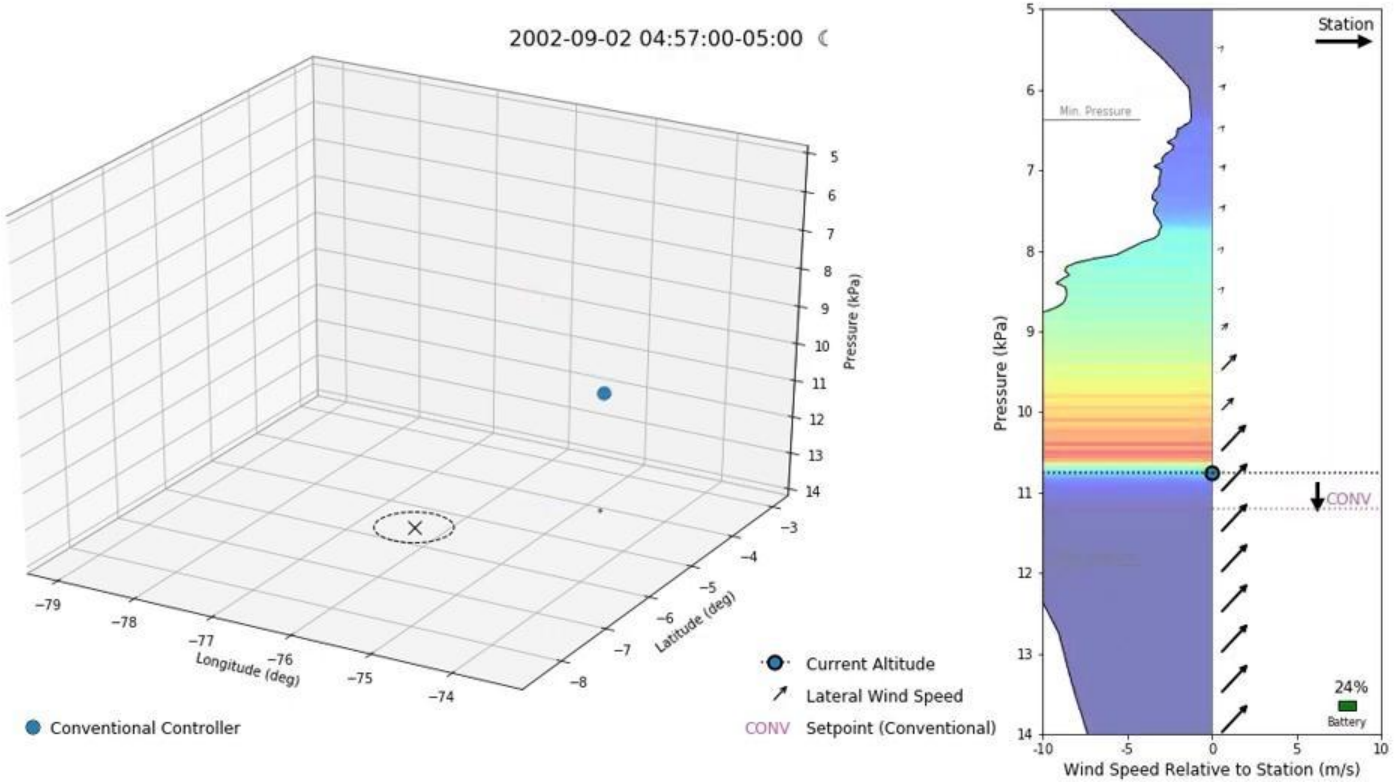

Distance to station (km, Deciles 1-9)
Simulation time (hours)

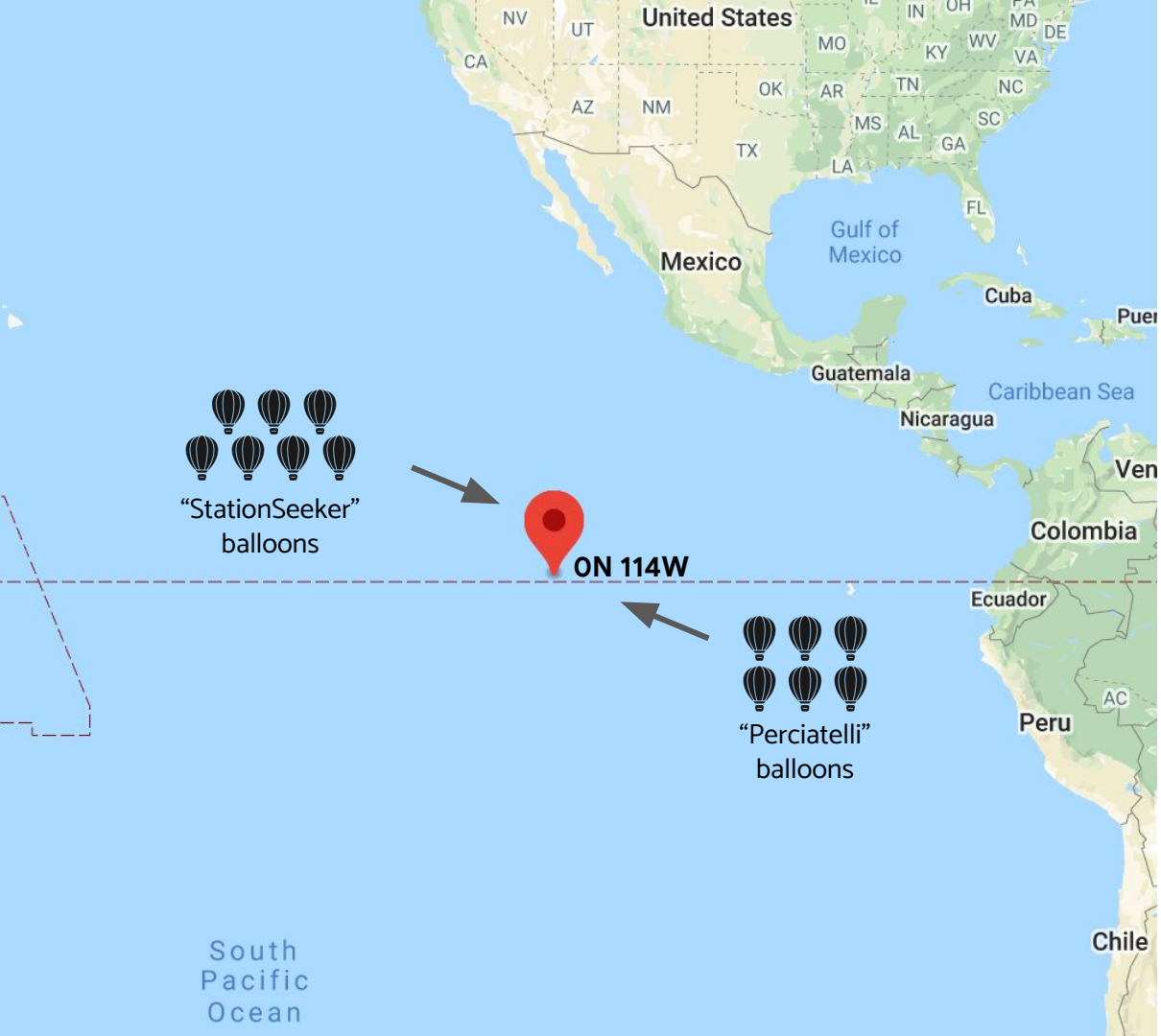

Learned controller
Conventional controller
Pairwise distances (km)
Simulation time (hours)


Tuned search
Learned (24 days)
7 days
2 days
1 day
12h
OPD (equal power)
StationSeeker
OPD controllers
StationSeeker-type controllers
Learned controller
TWR50 (%)
Power (W)
60
55
50
45
40
35
30
25
20
50
45
40
35
30
25
20
15


Perfect information limit
Plausible range of
max. benchmark TWR50
TWR50 (%)
70
60
50
40
30
20
10
0
55.1
40.5
51.6
Simulation
baselines
Learned
S. Seeker
OPD
Random
Passive


0 h
3 h
6 h
9 h
Probability Density
Probability Density
0
25
50
75
100
125
150
Return
Return
Descend
Stay
Ascend

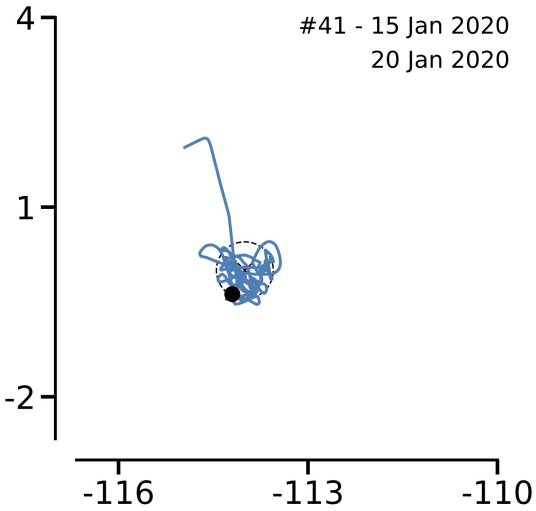

2002-09-02 04:57:00-05:00
Pressure (kPa)
Longitude (deg)
Latitude (deg)
Conventional Controller
Current Altitude
Lateral Wind Speed
CONV Setpoint (Conventional)
Station
Min. Pressure
CONV
24%
Battery
Wind Speed Relative to Station (m/s)



Pacific Ocean Experiment

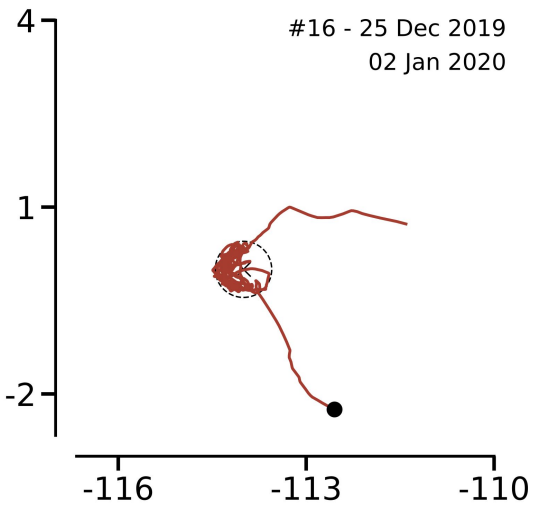
26 Oct 2019 – 25 Jan 2020
13 balloons
Total **2884** RL flight hours
Longest RL flight **~16 days**

StationSeeker

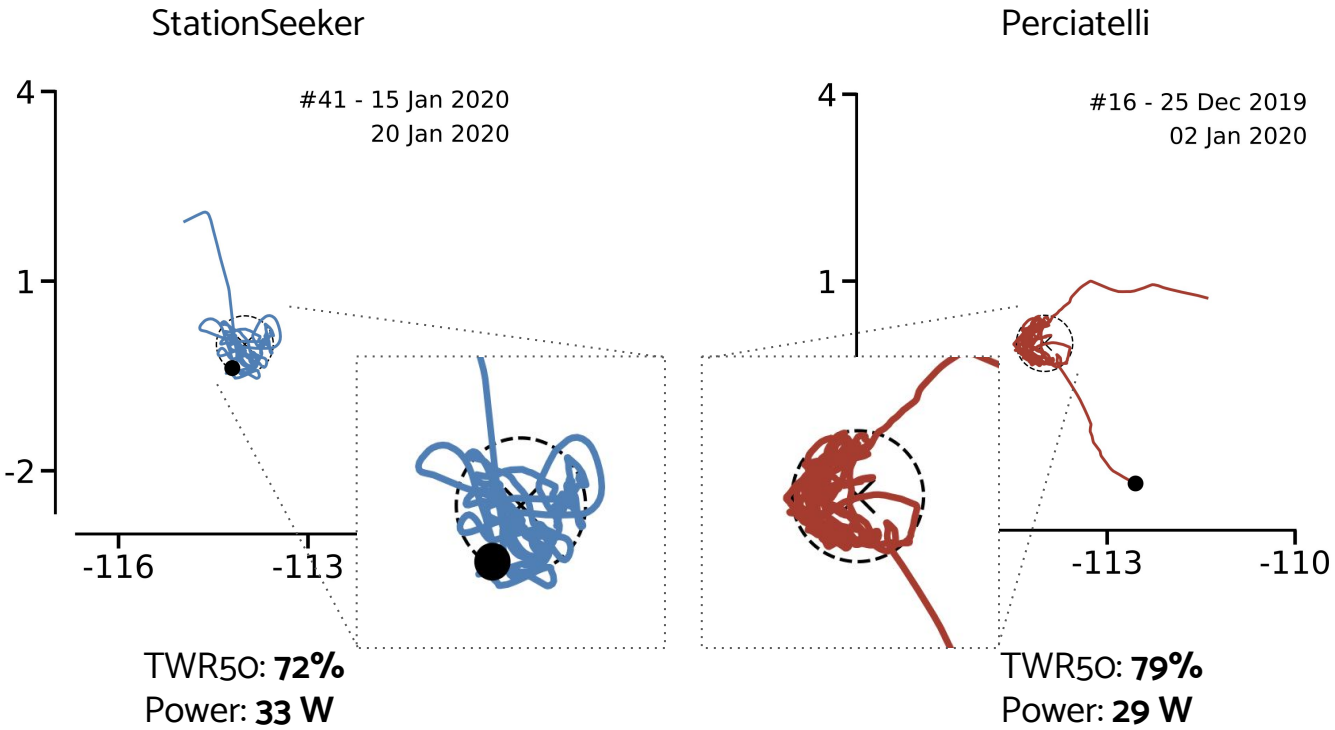


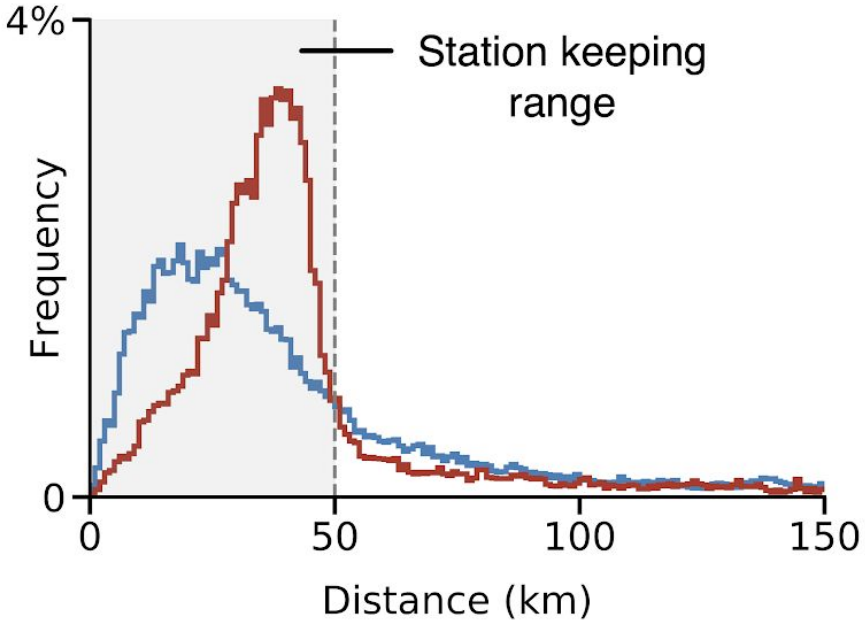
TWR50: **72%**
Power: **33 W**

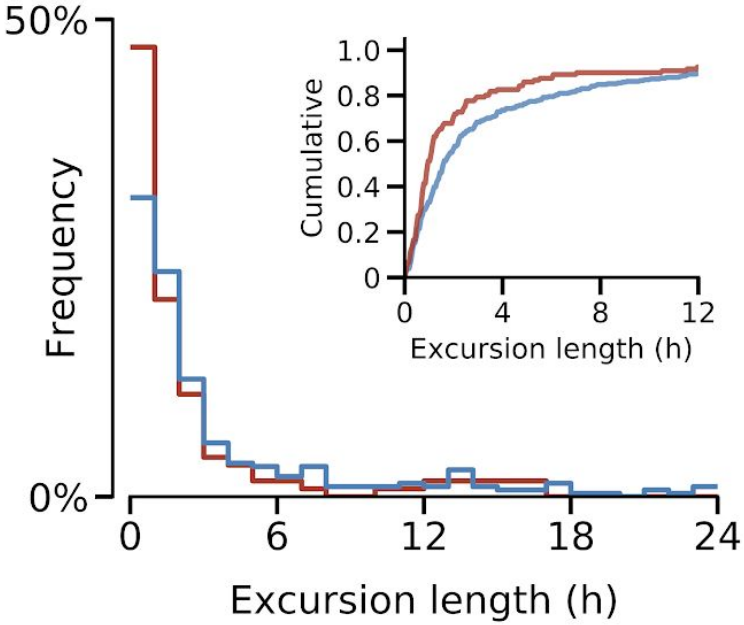
Perciatelli

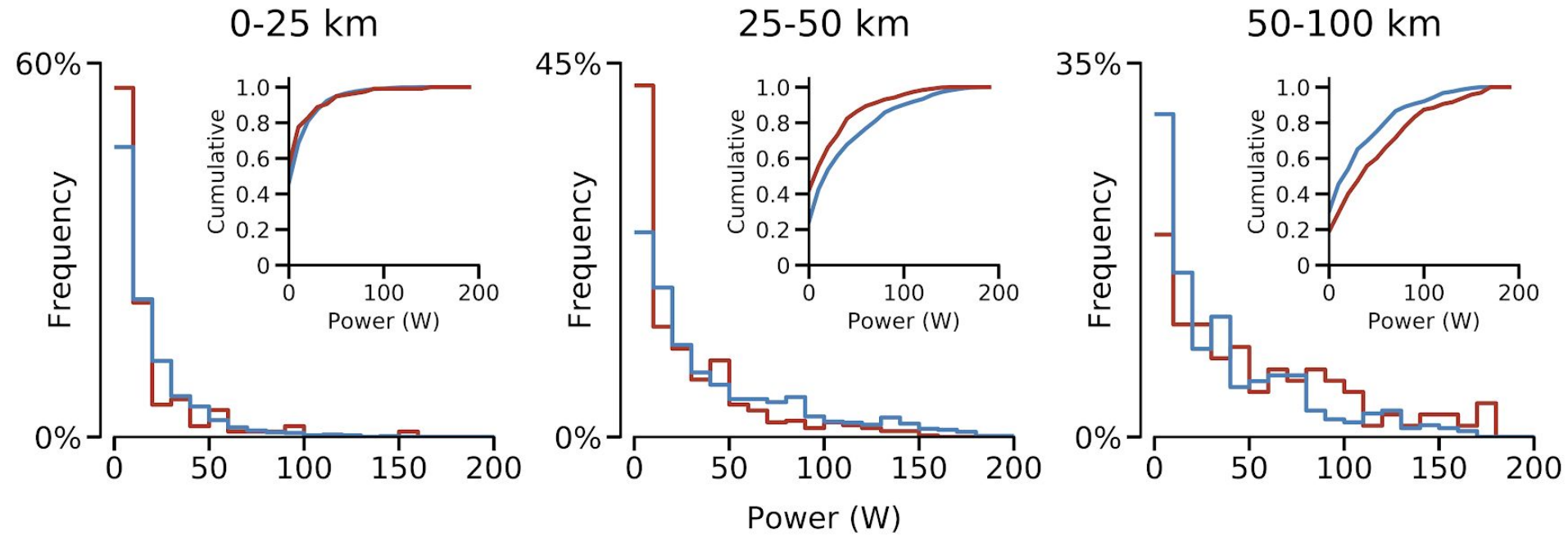


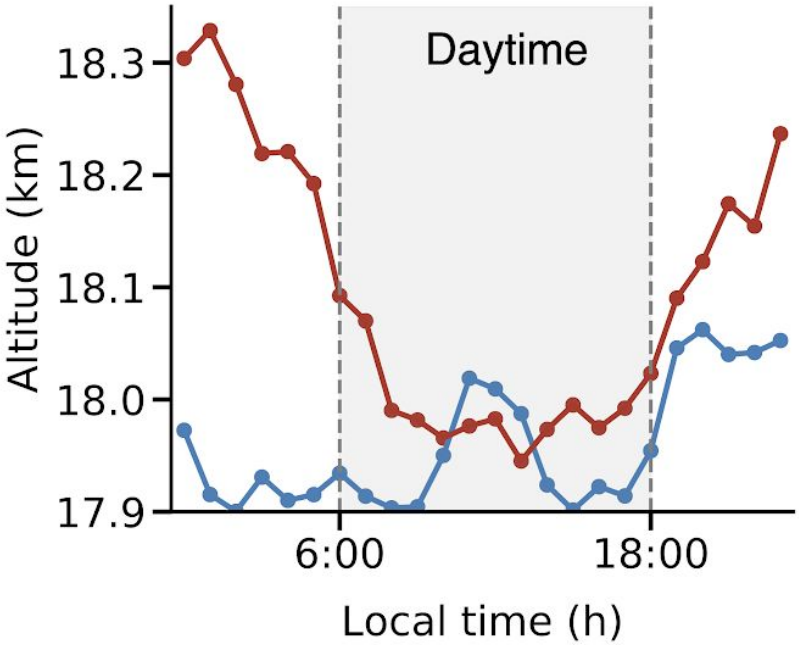
TWR50: **79%**
Power: **29 W**

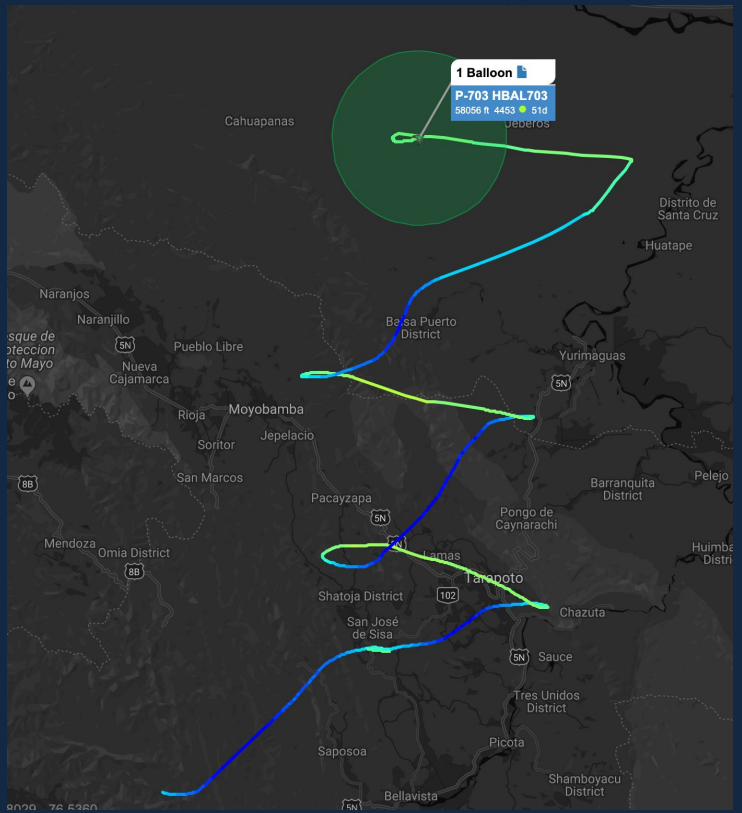

StationSeeker
#41 - 15 Jan 2020
20 Jan 2020
4
1
-2
-116
-113
TWR50: 72%
Power: 33 W
Perciatelli
#16 - 25 Dec 2019
02 Jan 2020
4
1
-113
-110
TWR50: 79%
Power: 29 W

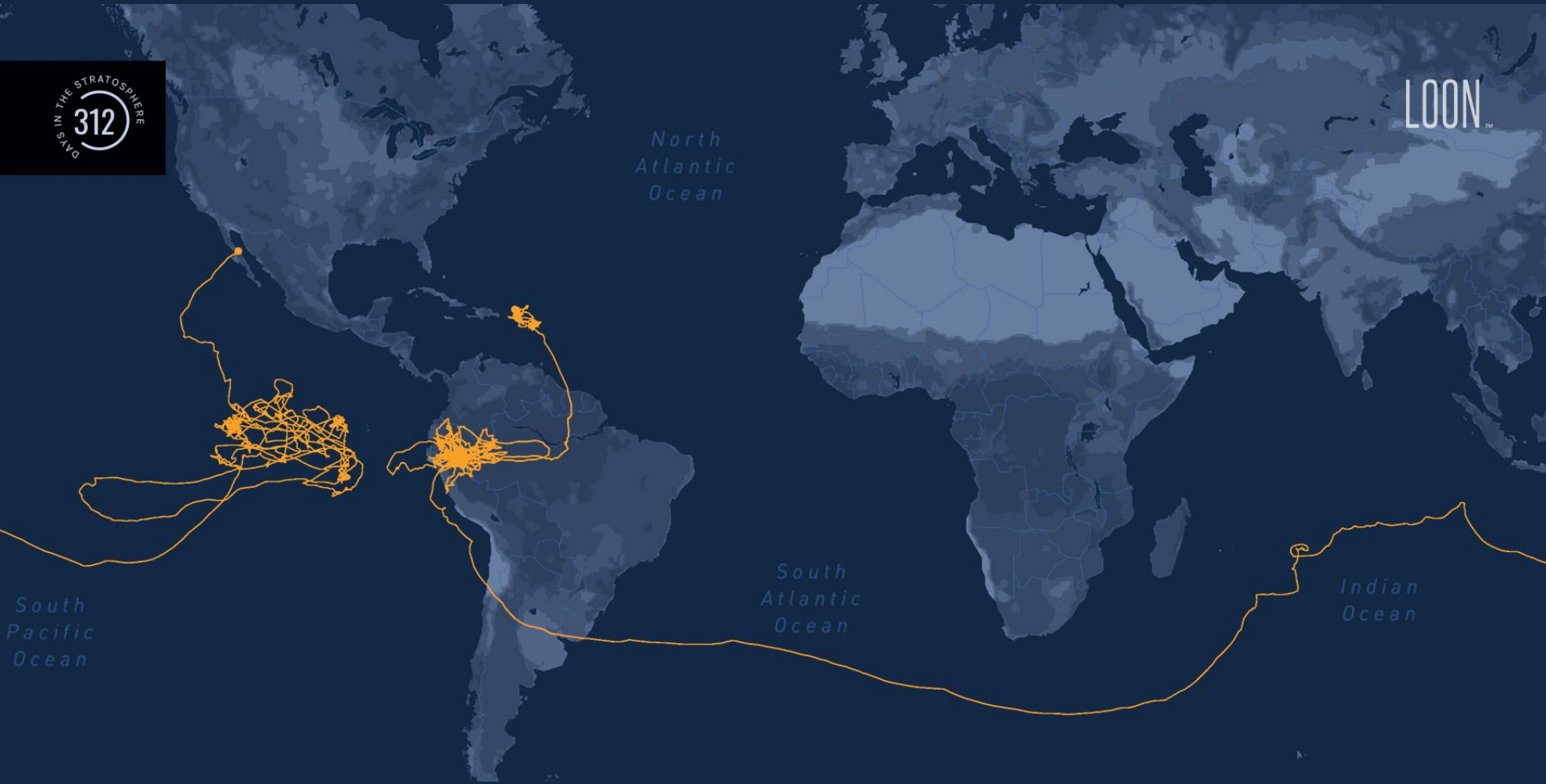

4%
Station keeping range
Frequency
0
0
50
100
150
Distance (km)


50%
Frequency
0%
0
6
12
18
24
Excursion length (h)
Cumulative
1.0
0.8
0.6
0.4
0.2
0
0
4
8
12
Excursion length (h)

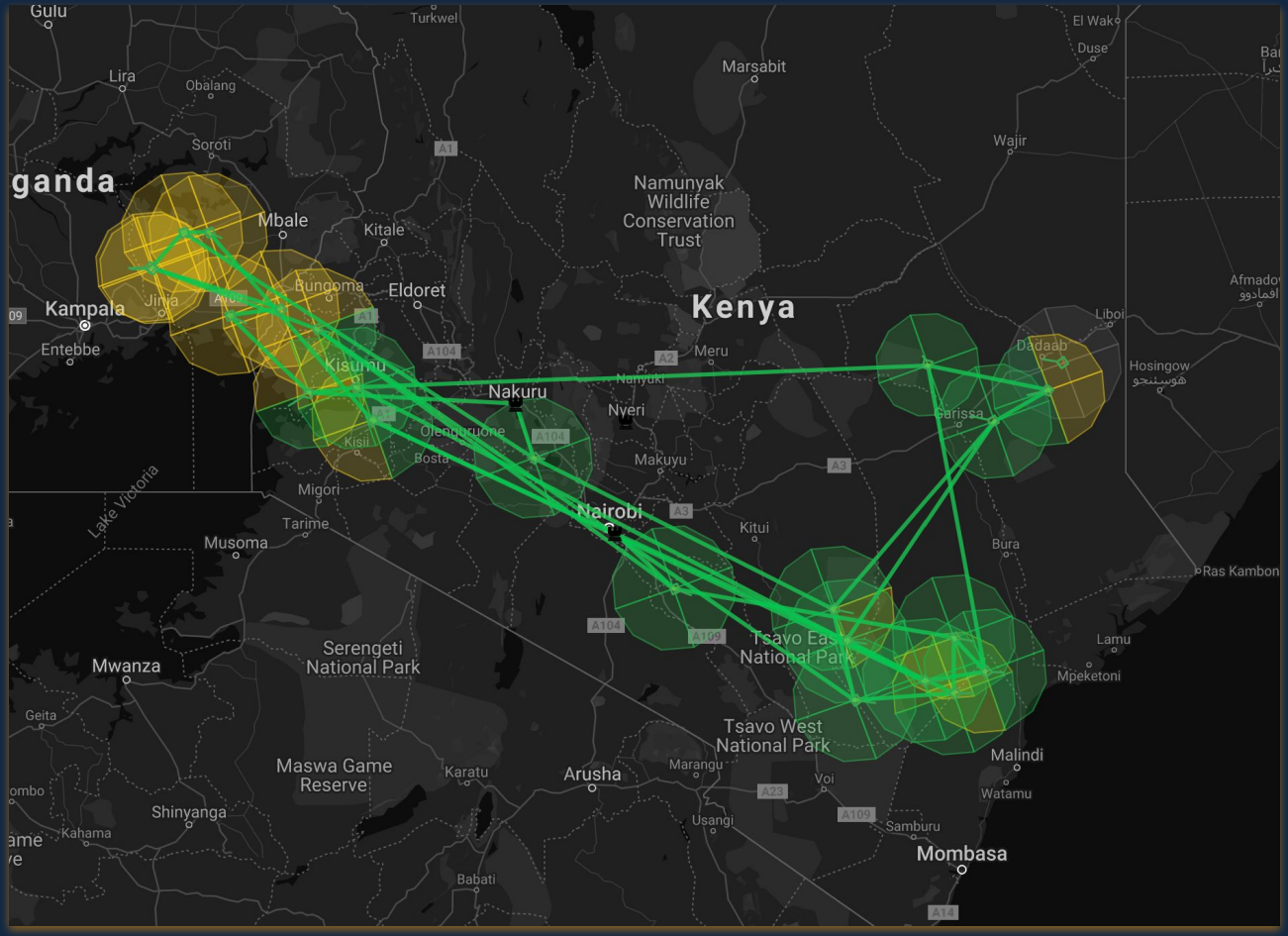

0-25 km
25-50 km
50-100 km
60%
45%
35%
0%
Frequency
Cumulative
1.0
0.8
0.6
0.4
0.2
0
0
50
100
150
200
Power (W)

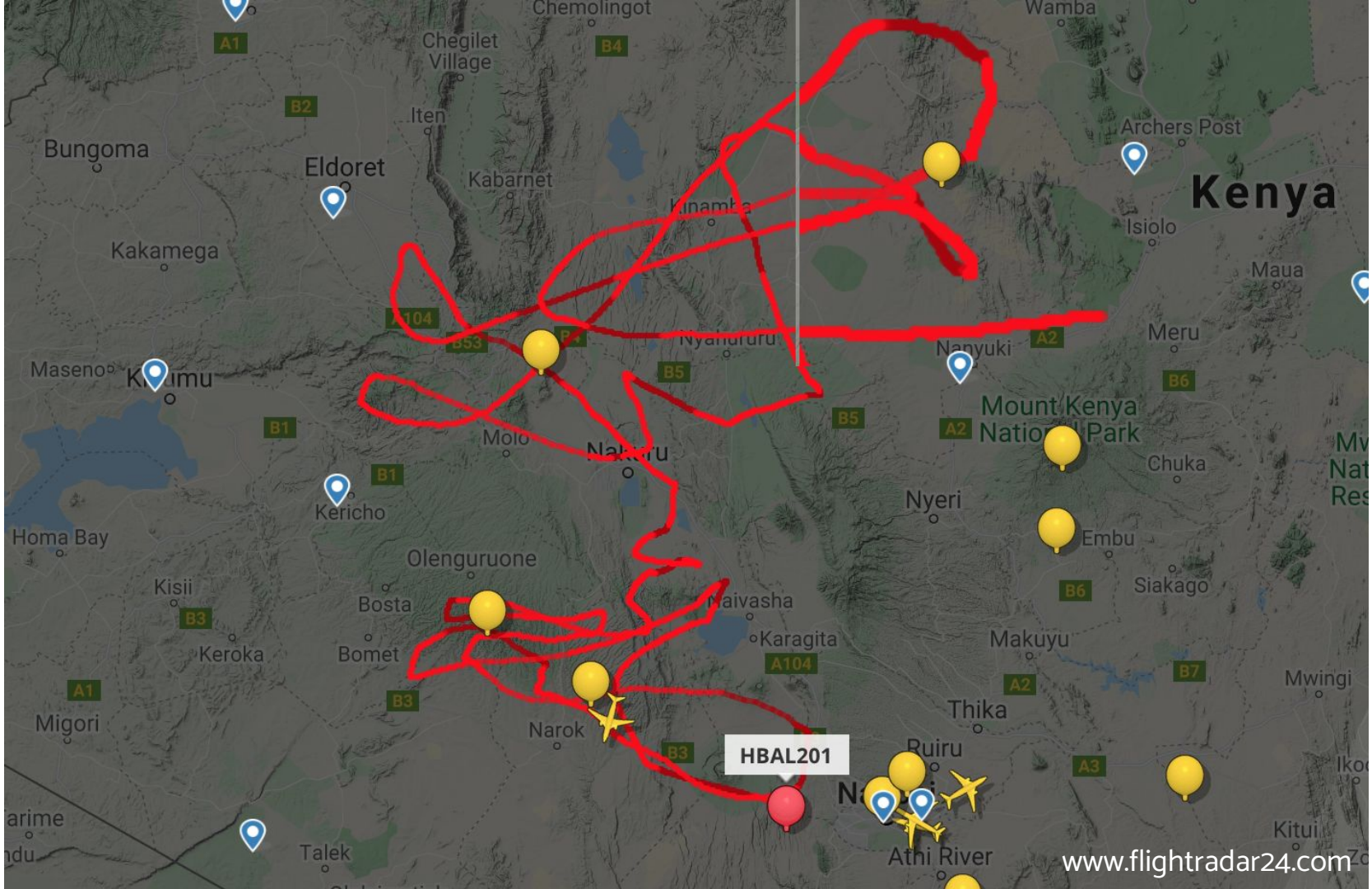

Daytime
18.3
18.2
18.1
18.0
17.9
Altitude (km)
6:00
18:00
Local time (h)





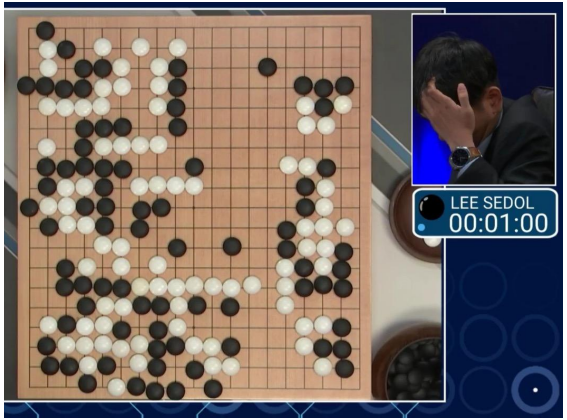
312 Days in the Stratosphere, Loon, Oct 28 2020.




Chemolingot
Chegilet Village
Wamba
Archers Post
Iten
Bungoma
Eldoret
Kabarnet
Kenya
Isiolo
Kakamega
Maua
Nyahururu
Nanyuki
Meru
Maseno
Kisumu
Mount Kenya National Park
Molo
Nakuru
Chuka
Kericho
Nyeri
Homa Bay
Embu
Olenguruone
Kisii
Bosta
Naivasha
Siakago
Karagita
Keroka
Bomet
Makuyu
Migori
Narok
Thika
HBAL201
Ruiru
Talek
Athi River
Kitui
www.flightradar24.com

Why does this matter?

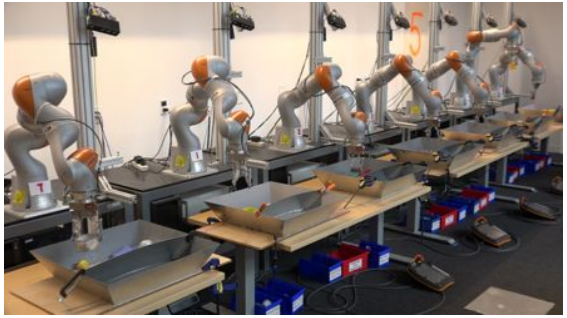
Deep reinforcement learning



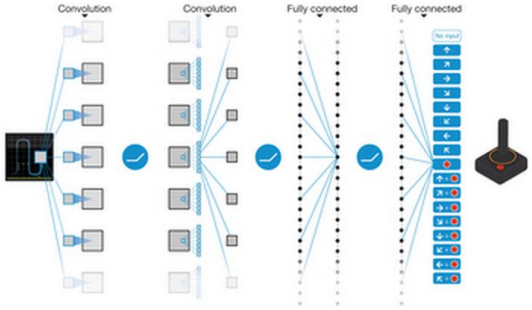
Silver, Huang, Maddison, et al. (2016, 2017)

Bellemare, Naddaf, Veness, Bowling (2013)

Bard, Foerster, Chandar, et al. (2020)



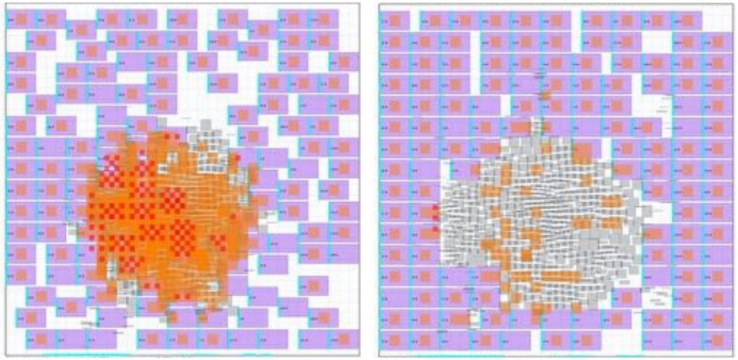
Levine et al., (2016)
Kalashnikov, Irpan, et al. (2018)



Mnih, Silver, Kavukcuoglu, Rusu, Veness, Bellemare, et al. (*Nature*, 2015)

Vinyals et al. (2019)

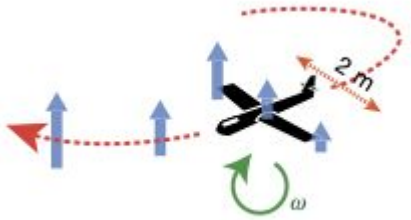
OpenAI et al. (2019)



Mirhoseini, Goldie, et al. (arXiv, 2020)

Managing Fragmented Fire-Threatened Landscapes with Spatial Externalities

Christopher J. Lauer, Claire A. Montgomery, and Thomas G. Dietterich



Reddy et al. (2018)

Won et al., 2020

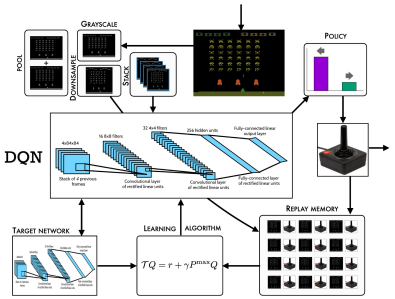
Glavic et al., 2017

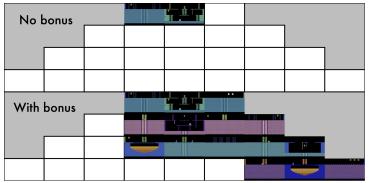
YouTube

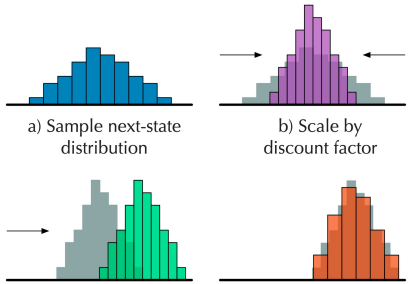
Ie et al., 2019

Challenging our paradigms

The Arcade Learning Environment, Bellemare, Naddaf, Veness, Bowling (2013).







"Real-world applications drive fundamental research questions"

– Chris Bishop

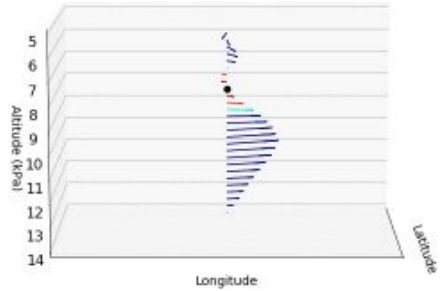
Decisions from data
Controlling complex systems with reinforcement learning

Marc G. Bellemare
Google Brain, Montreal; Mila

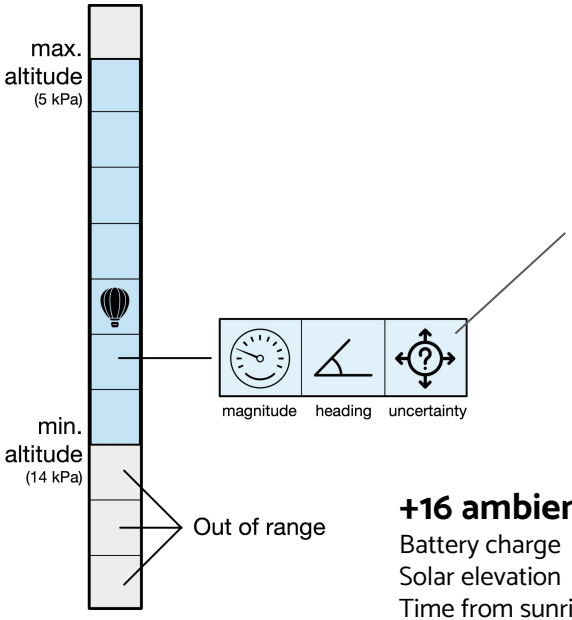
With: Salvatore Candido, Pablo Samuel Castro, Jun Gong, Marlos C. Machado, Subhodeep Moitra, Sameera S. Ponda, Ziyu Wang

IVADO

RL for balloon navigation: inputs



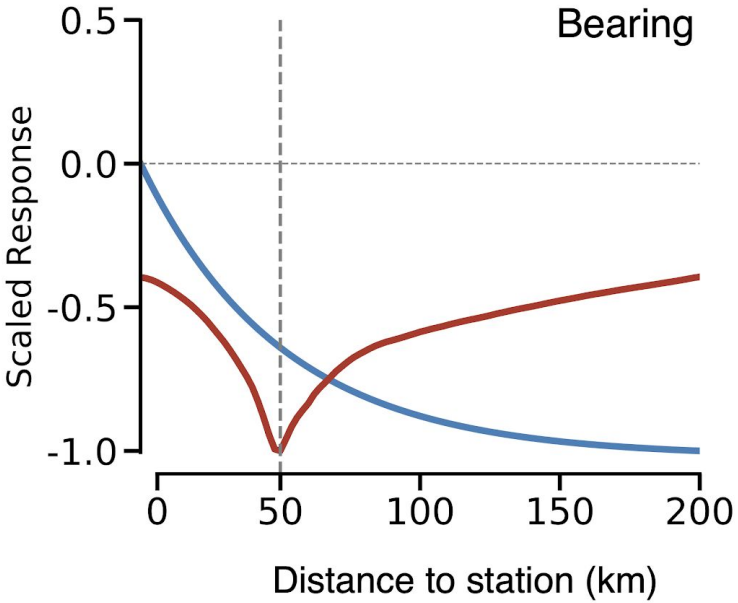
Forecast + measurements + Gaussian process = **wind column**

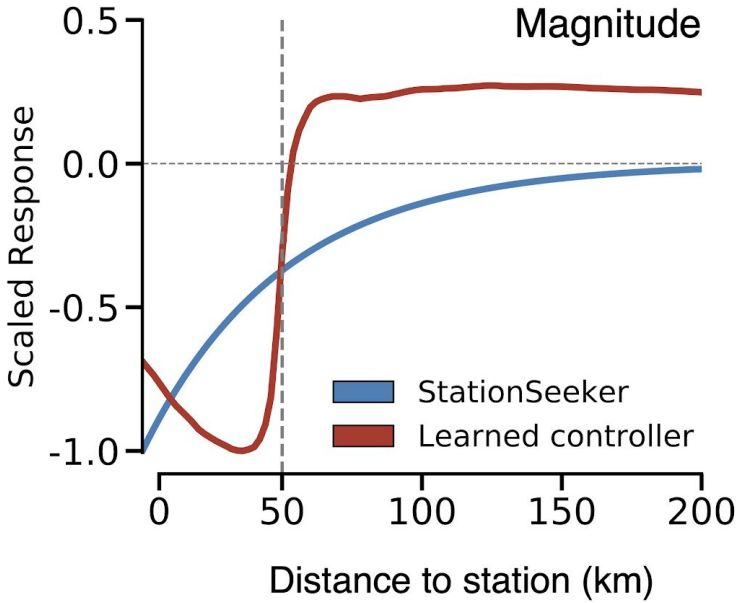


The uncertainty term at each altitude level acts as a **belief feature**

+16 ambient variables:

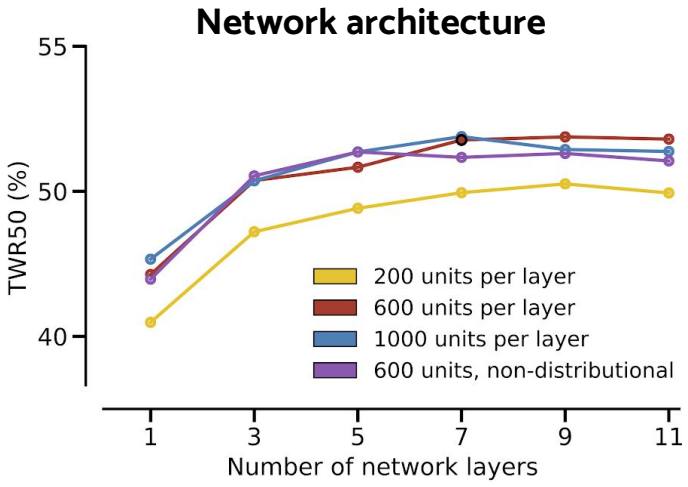
Battery charge
Solar elevation
Time from sunrise
Distance to station
Internal pressure ratio
etc.

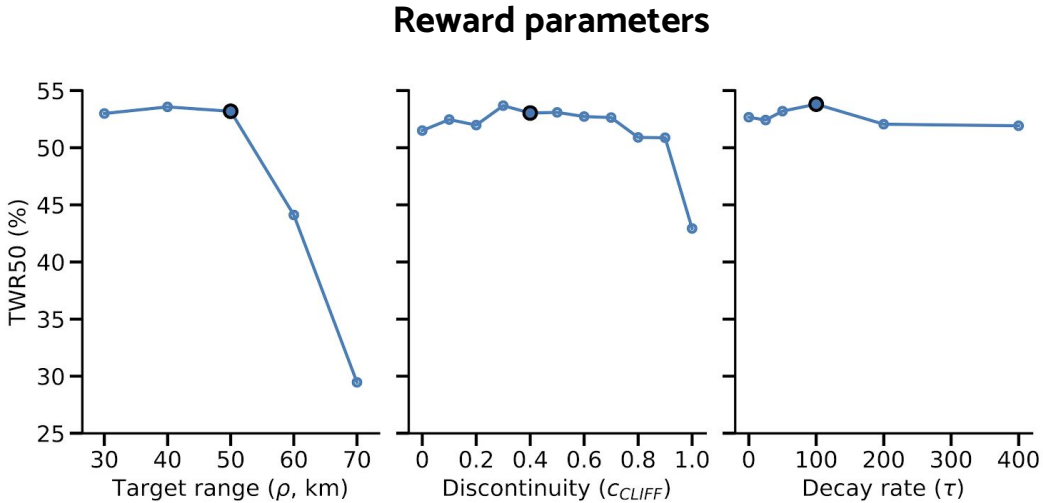


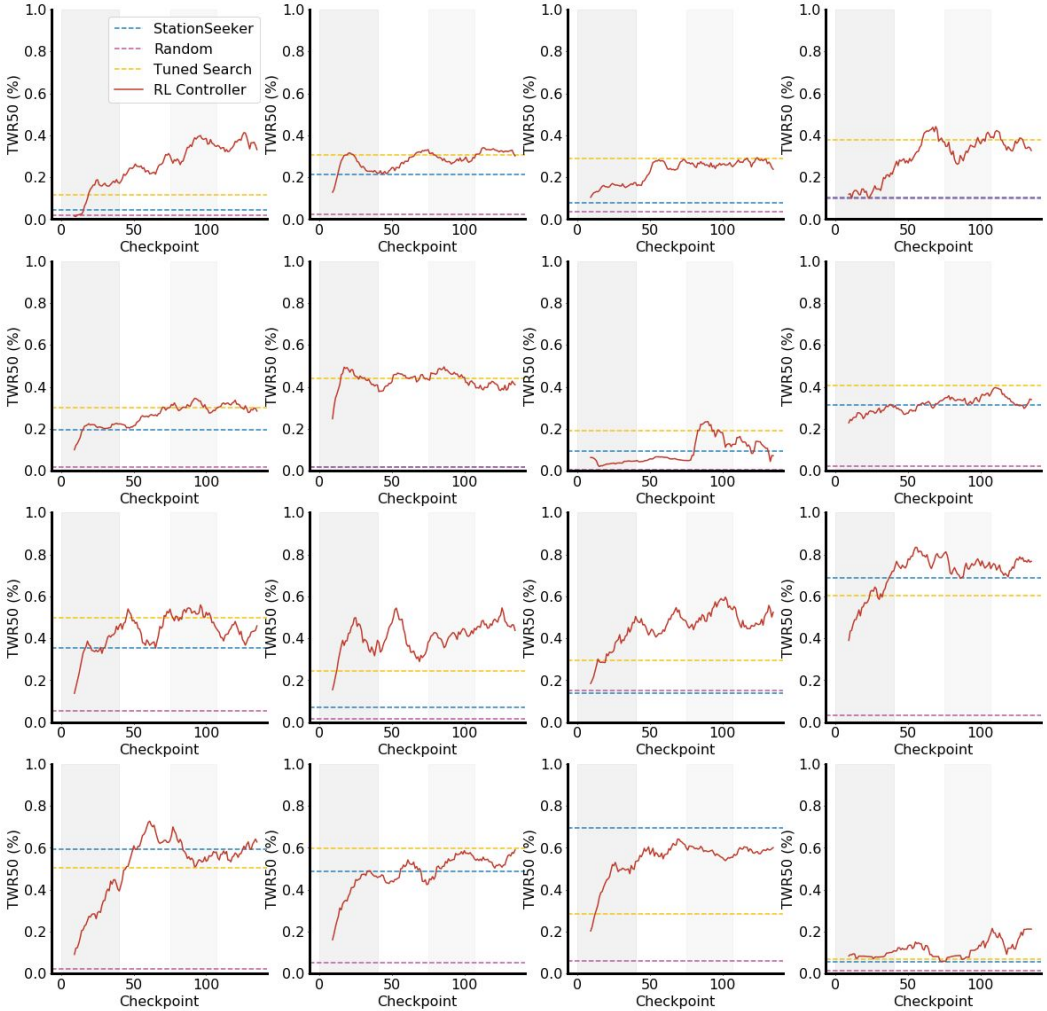


Sensitivity analysis

Twelve 2-day flights
Estimate **gradient of value function** w.r.t. wind characteristics (finite differences)
Plot average gradient **as a function of distance**


Network architecture
55
50
40
TWR50 (%)
200 units per layer
600 units per layer
1000 units per layer
600 units, non-distributional
1
3
5
7
9
11
Number of network layers


Reward parameters
55
50
45
40
35
30
25
TWR50 (%)
30
40
50
60
70
Target range (ρ, km)
0
0.2
0.4
0.6
0.8
1.0
Discontinuity (c_{CLIFF})
0
100
200
300
400
Decay rate (τ)


StationSeeker
Random
Tuned Search
RL Controller
TWR50 (%)
Checkpoint
0.0
0.2
0.4
0.6
0.8
1.0
0
50
100