Guiding drones by information gain

Alouette van Hove*[1], Kristoffer Aalstad[1], and Norbert Pirk[1]

[1]Department of Geosciences, University of Oslo, Norway
a.van.hove@geo.uio.no

Abstract

The accurate estimation of locations and emission rates of gas sources is crucial across various domains, including environmental monitoring and greenhouse gas emission analysis. This study investigates two drone sampling strategies for inferring source term parameters of gas plumes from atmospheric measurements. Both strategies are guided by the goal of maximizing information gain attained from observations at sequential locations. Our research compares the myopic approach of infotaxis to a far-sighted navigation strategy trained through deep reinforcement learning. We demonstrate the superior performance of deep reinforcement learning over infotaxis in environments with non-isotropic gas plumes.

1 Introduction

As a result of global warming, permafrost soils are thawing and will potentially release large amounts of greenhouse gases such as methane and carbon dioxide into the atmosphere [1]. Understanding the scale and distribution of these greenhouse gas sources helps improve the reliability of climate projections.

Determining the location and flux (i.e. emission rate) of greenhouse gas sources remains challenging as they cannot be observed directly. Instead, these source term parameters must be inferred through the analysis of observational data, such as gas concentrations and other atmospheric variables. This inverse problem is known as source term estimation (STE) [2]. The Bayesian inversion approach combines observational data with an atmospheric plume model, accounting for uncertainties in both, to yield a probability distribution over possible source term parameters. Sequentially applying Bayes' theorem to each observation updates the parameters' probability distribution [3] so as to constrain uncertainty in the source term estimates.

Observations can be acquired using static sensors (e.g. measurement towers), mobile sensors (e.g. drones), or a combination of the two. Although mobile sensors offer the advantage of flexibility and adaptability, the challenges of devising the optimal sampling strategy become even more apparent. Sample trajectories can be pre-planned in an offline manner, for example, by selecting a sweeping "lawnmower" pattern, or they can be adapted to the environment in an online manner [4]. The latter is known as informative path planning (IPP).

In this paper, we explore two approaches to IPP for addressing the STE problem. The first approach involves a modified version of infotaxis, an online search strategy initially introduced by Vergassola et al. [5] for source localisation. Our work extends this algorithm to estimate multiple source term parameters. The second approach uses deep reinforcement learning (DRL) [6] to develop a sampling strategy. We train a mobile drone-based sensor to obtain the most certain estimates of the source term parameters within a finite number of observations, simulating the constraints of a drone's limited battery life. This study aims to determine whether a DRL strategy with an information-based reward can outperform the infotaxis strategy in the context of STE.

Recent studies have explored the application of DRL for guiding mobile sensors in source seeking tasks [7–11]. However, these studies focused solely on source localisation, neglecting source strength estimation. Moreover, task completion in [7, 9–11] was determined based on the proximity of the mobile sensor to the source. This is impractical for the real-world identification of unknown greenhouse gas sources whose location is uncertain. This study seeks to jointly address both the source localisation and strength problem. While Park et al. [12] did consider the STE problem, they aimed to minimize the number of observations needed for accurate STE, rather than minimizing uncertainty within a single sampling flight. In line with van Hove et al. [13], we apply an information-based reward function. Whereas this earlier study estimated the source strength at a known location using tabular RL, we apply deep RL herein to account for the larger parameter space.

The study's primary contributions are as follows: (a) Application of DRL to address the STE problem under pertinent constraints for greenhouse gas mapping in the field - considering unknown source locations, fixed flight time due to battery limitations, and noisy observations by the sensor. (b) Comparing the performance of DRL with an information-based reward to infotaxis. (c) Comparison between two neural network architectures: a feed forward structure with fully connected layers (as utilized by [7–

*Corresponding Author.

Proceedings of the 5th Northern Lights Deep Learning Conference (NLDL), PMLR 233, 2024.
2024 Alouette van Hove, Kristoffer Aalstad, & Norbert Pirk. This is an open access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/4.0/).

11]) and a convolutional neural network.

2 Problem description

The navigation of mobile sensors for STE is a sequential decision-making problem characterized by uncertainty and partial observability. The mobile sensor cannot directly perceive the true state of the emissions but instead receives incomplete and uncertain observations related to gas concentration. In a mathematical framework, the problem can be represented as a Partially Observable Markov Decision Process (POMDP) [14], that can be framed as a belief Markov Decision Process (belief-MDP) where states are replaced by belief states. In this context, the mobile sensor - *the agent* - maintains a probability distribution over the unknown source term parameters - *its belief* - while navigating a two-dimensional grid. At each time step, the agent can choose from a set of five actions: $\{\rightarrow, \leftarrow, \downarrow, \uparrow, \text{stay}\}$. After executing an action, the agent receives a noisy observation which it uses to update its beliefs about the source term parameters through Bayesian inference. Solving the navigation problem for STE with mobile sensors means finding the optimal strategy - *the policy* - for mapping belief states to actions. The agent's objective is to maximize a reward function that incentivizes it to select actions leading to more certain source term estimations.

2.1 Observation model

A greenhouse gas source emits detectable particles that are dispersed through domain $\mathcal{A}$ under turbulent transport conditions. We use the advection-diffusion plume model presented in [5] to compute the mean number of detected particles, or "hits", by an agent at position $\mathbf{x} = [x, y] \in \mathcal{A}$ during a sampling time interval Δt. The source term parameter vector $\boldsymbol{\theta}$ includes the source location $\mathbf{x}_\text{s} = [x_\text{s}, y_\text{s}] \in \mathcal{A}$ and its source strength (i.e. flux) $\phi \in \mathbb{R}^+$. In two dimensions, the mean number of hits detected by an agent with body radius r is given by Eq. (1)

$$\mu(\boldsymbol{\theta}, \mathbf{x}) = \phi \Delta t \frac{1}{\ln(\frac{\lambda}{r})} \exp\left(\frac{-(y - y_\text{s})V}{2D}\right) K_0\left(\frac{\|\mathbf{x} - \mathbf{x}_\text{s}\|}{\lambda}\right), \tag{1}$$

where the wind has been taken to blow in the negative y-direction. K_0 is the modified Bessel function of the second kind of order 0. In this model, the gas particles have a finite lifetime τ. Mean wind speed V and effective diffusivity D (i.e. turbulent and molecular diffusivity) affect the dispersion length scale λ

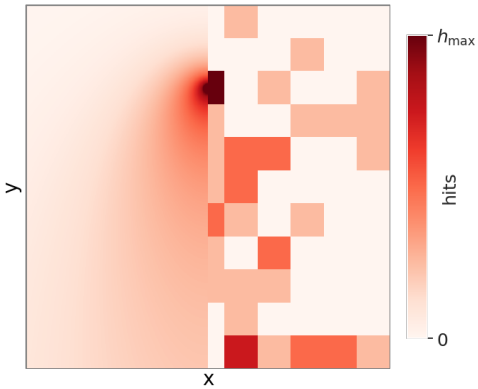

Figure 1. Example of a hit map: (left half) map of the mean number of hits μ calculated by the analytic plume model Eq. (1), and (right half) a possible map of noisy sensor measurements h from Eq. (3).

of the particles by Eq. (2)

$$\lambda = \sqrt{\frac{D\tau}{1 + \frac{V^2\tau}{4D}}}\,. \tag{2}$$

An example of a mean hit map is shown on the left half of Figure 1. The maximum number of hits is fixed to an upper bound, h_max. The mean hit map becomes isotropic in the absence of wind.

The stochastic data generating process for the number of hits detected by the sensor during sampling interval Δt is modelled by a Poisson distribution. The probability that the sensor detects h hits is given by Eq. (3)

$$\Pr(h|\mu) = \frac{\mu^h}{h!} \exp(-\mu)\,. \tag{3}$$

The right half of Figure 1 shows a possible measurement map. The difference between the left half and right half of the map demonstrates the challenge of STE in turbulent conditions. Observations are sparse and noisy, and gradient-based sampling strategies for STE are therefore often not optimal and can fail in turbulent flows.

2.2 Bayesian inference

Each hit observation provides information about the source term parameters. The agent applies Bayesian inference after each observation to maintain an up-to-date belief about the source term parameters $\Pr(\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a discrete random variable of all possible source locations and fluxes. In particular, Bayes' rule (4) maps the prior belief $\Pr(\boldsymbol{\theta}_n)$ to the posterior belief $\Pr(\boldsymbol{\theta}_n|h_n)$ after a hit observation h_n at time t_n

$$\Pr(\boldsymbol{\theta}_n|h_n) = \frac{\Pr(h_n|\boldsymbol{\theta}_n)\Pr(\boldsymbol{\theta}_n)}{\Pr(h_n)}\,, \tag{4}$$

where the likelihood $\Pr(h_n|\boldsymbol{\theta}_n)$ is computed by Eq. (3) with μ given by Eq. (1). All probabilities are implicitly conditioned on the plume model and background information, including agent's location $\mathbf{x}$ and wind speed V, among others. The evidence $\Pr(h_n) = \sum_i \Pr(h_n|\boldsymbol{\theta}_n^{(i)}) \Pr(\boldsymbol{\theta}_n^{(i)})$ acts as a normalizing constant such that the posterior belief over all possible $\boldsymbol{\theta}_n^{(i)}$ sums to one. Our agent applies Eq. (4) sequentially in time by letting the posterior at time t_n become the prior at time t_{n+1}, $\Pr(\boldsymbol{\theta}_{n+1}) = \Pr(\boldsymbol{\theta}_n|h_n)$, resulting in dynamic belief updating.

3 Sampling strategies

Within the belief-MDP framework, the agent's belief state is denoted as $s = (\mathbf{x}, \Pr(\boldsymbol{\theta}))$. In this study, both DRL and infotaxis policies utilize a reward function based on information entropy. Information entropy H [15] is non-negative and measures the uncertainty of a probability distribution. The information entropy (measured in nats) of a belief state s is given by Eq. (5)

$$H(s) = -\sum_i \Pr(\boldsymbol{\theta}^{(i)}) \ln \Pr(\boldsymbol{\theta}^{(i)}) \,. \tag{5}$$

Note that information entropy of a belief state is not directly affected by the agent's location $\mathbf{x}$, and relies only on $\Pr(\boldsymbol{\theta})$, which in turn depends on the agent's sampling strategy and observations. The initial prior belief of the agent is a uniform distribution over all possible source locations and fluxes, resulting in maximum information entropy. Conversely, the information entropy of a Kronecker delta distribution, where the probability is one for a specific $\boldsymbol{\theta}^{(i)}$ and zero for all others, attains the minimum information entropy of zero.

3.1 Infotaxis

Infotaxis [5] is a greedy policy which sequentially maximizes its reward function. At each step, the agent selects an action a that maximizes the expected information gain G [9], such that Eq. (6) follows

$$\pi_{\text{infotaxis}}(s) = \arg\max_a G(s,a) \,. \tag{6}$$

Expected information gain is the expected difference in information entropy between the agent's current belief state s and all possible successor belief states s', given by Eq. (7)

$$G(s,a) = H(s) - \sum_{s'} \Pr(s'|s,a) H(s') \,. \tag{7}$$

The agent can reach different successor belief states depending on the hits it receives after performing action a. Combining Eq. (6) and (7) gives Eq. (8)

$$\pi_{\text{infotaxis}}(s) = \arg\min_a \sum_{s'} \Pr(s'|s,a) H(s') \,. \tag{8}$$

such that the policy can also be formulated as selecting the action that transitions the agent to the successor belief state with the minimum expected information entropy (i.e. minimum uncertainty).

3.2 Deep reinforcement learning

Infotaxis is a myopic approach, aiming to maximize the expected reward of the next action. In contrast, RL trains a policy to maximize the cumulative reward obtained throughout an entire *episode*, considering a sequence of actions. The essence of RL lies in identifying the optimal policy associated with the optimal value function $v_*(s)$. This represents the maximum expected *cumulative* reward that can be achieved from a given belief state. The large number of possible beliefs $\Pr(\boldsymbol{\theta})$ in a belief state s prevents the use of exact methods for obtaining $v_*(s)$. Consequently, we approximate $v_*(s)$ through a parameterized function $\hat{v}(s, \mathbf{w})$, in the form of a deep neural network [16] with weight vector $\mathbf{w}$. The Bellman optimality equation for the approximate value function [6] is given by Eq. (9)

$$\hat{v}(s, \mathbf{w}_*) = \max_a \sum_{s',r} \Pr(s', r|s,a) \left[r + \hat{v}(s', \mathbf{w}_*) \right] \,, \tag{9}$$

where r is the reward. We use a reward function based on information gain (7): $r = -H(s')$. This reward function depends only on successor state s', such that Eq. (10) follows as the Bellman optimality equation for this study

$$\hat{v}(s, \mathbf{w}_*) = \min_a \sum_{s'} \Pr(s'|s,a) \left[H(s') + \hat{v}(s', \mathbf{w}_*) \right] \,. \tag{10}$$

Solving for the optimal weights $\mathbf{w}_*$ involves minimizing the residual error, known as the Bellman optimality error (11) [9]

$$L(\mathbf{w}) = \mathbb{E}_s \left[\min_a \sum_{s'} \Pr(s'|s,a) \left[H(s') + \hat{v}(s', \mathbf{w}) \right] - \hat{v}(s, \mathbf{w}) \right]^2 \,. \tag{11}$$

The expectation is taken over belief states s visited when following policy $\hat{\pi}$ derived from $\hat{v}$ and defined by Eq. (12)

$$\hat{\pi}(s, \mathbf{w}) = \arg\min_a \sum_{s'} \Pr(s'|s,a) \left[H(s') + \hat{v}(s', \mathbf{w}) \right] \,. \tag{12}$$

The functional $L(\mathbf{w})$ is the *loss function*, and *training* the neural network entails optimizing the weights $\mathbf{w}_\star = \arg\min_{\mathbf{w}} L(\mathbf{w})$ using stochastic gradient descent, aiming to minimize the loss function. The subsequent DRL policy is given by Eq. (12) evaluated with the optimized weights $\pi_{\mathrm{DRL}}(s, \mathbf{w}_*) = \hat{\pi}(s, \mathbf{w}_\star)$. Note that by initializing the weights of the value function in proximity to zero, the untrained DRL policy (12) closely resembles the infotaxis policy (8).

Our DRL algorithm is based on Deep Q-Networks (DQN) [17], an extension of Q-learning to deep neural networks. While most RL algorithms are model-free and use state-action values $q(s, a)$, our study is inspired by the approach of Loisy and Eloy [9] and Loisy and Heinonen [11] (with accompanying code [18]), and adopts a model-based approach due to our exact knowledge of transition probabilities $\Pr(s', r|s, a)$. This allows us to work with $\hat{v}(s, \mathbf{w})$ instead of $\hat{q}(s, a, \mathbf{w})$.

We utilize a feedforward neural network and compare two different network architectures: (a) an architecture with 3 fully connected hidden layers each containing 32 units, and (b) an architecture with 4 hidden convolutional layers with a filter size of 32 and a kernel size of 3. Average pooling with kernel size 2 and stride 2 is applied. Both network architectures use rectified linear units (ReLU) for activations, and a linear output layer. The network sizes are tuned such that adding more weights does not increase their performance any further. The weight parameters are trained over 20,000 episodes using stochastic gradient descent with mini-batches of size 128.

During preprocessing, we transform belief state $s = (\mathbf{x}, \Pr(\boldsymbol{\theta}))$ containing the agent's position and source term probability distribution ($N_x \times N_y \times N_\phi$ tensor), into a source term probability distribution centered on the agent ($(2N_x-1)\times(2N_y-1)\times N_\phi$ tensor) [18]. Tensor elements outside the physical domain are set to zero. When used as an input for architecture (b), the tensor can be conceived as an image representing spatial data with dimensions $(2N_x - 1) \times (2N_y - 1)$ pixels and N_ϕ channels. The tensor is flattened when used as an input for architecture (a).

4 Results

We adopt a non-dimensional representation of observation model (1) whereby hit counts are governed by dimensionless parameters: source flux $\tilde{\phi}$, location $\tilde{\mathbf{x}}$, agent radius $\tilde{r}$, gas lifetime $\tilde{\tau}$, mean wind speed $\tilde{V}$, and effective diffusivity $\tilde{D}$. We assume that the source is located within a domain of 11 x 11 grid cells and has one of the following fluxes: $\tilde{\phi} = \{1, 2, 3, 4, 5\}$. As a result, we consider 605 possible source term scenarios. The constants $\tilde{r} = 0.5$ and $\tilde{\tau} = 10,000,000$ are fixed. In sections 4.1 and 4.2, we maintain $\tilde{V} = 2$ and $\tilde{D} = 2$, while other values for $\tilde{V}$ and $\tilde{D}$ are explored in section 4.3. The agent starts in the center of the domain and performs 20 actions. The evaluation statistics presented herein are derived from 5,000 episodes.

4.1 Deep neural network

We assess the performance of a DRL policy approximated through a fully connected (FC) architecture and a convolutional neural network (CNN) as outlined in section 3.2. The FC architecture attains an average cumulative entropy of 66 nats, while the CNN performs slightly better with an average of 65 nats. In this study, we define *success* as the agent's ability to estimate the correct source location and flux with a final posterior probability of $\Pr \geq 0.5$. The FC architecture achieves success in 70% of episodes, while the CNN achieves this in 71% of episodes. The CNN architecture is used in subsequent sections of this study.

4.2 DRL versus infotaxis

We observe that the infotaxis policy achieves success in 58% of episodes (compared to 71% for DRL). Figure 2 shows a scenario with two agents navigating an environment with a gas source with $\tilde{\phi} = 2$ at $\tilde{\mathbf{x}}_\mathrm{s} = (9, 1)$: one agents follows the infotaxis policy, while the other adheres to the DRL policy. In the absence of receiving any hits, both agents move downwind (negative y-direction). The DRL agent follows a sweeping pattern downwind, whereas the infotaxis agent traverses along the edge of the domain. Upon detecting hits, the entropy of the agent's belief state reduces more significantly compared to zero hit events. Over the entire episode, the DRL agent receives more valuable information than the infotaxis agent, leading to a greater reduction in uncertainty of its belief state, and subsequently a superior (i.e. smaller) cumulative entropy.

We investigate whether DRL outperforms infotaxis across all source term scenarios considered in this study. As demonstrated in Table 1, the success rate increases with source strength, consistently favoring the DRL approach over infotaxis. By evaluating the performance across different source locations (not shown), we find that the DRL policy surpasses the success rate of infotaxis in 76% of source locations, primarily in the upper and right segments of the domain. Conversely, infotaxis outperforms DRL in 24% of source locations, mainly in the lower-left segment of the domain. As demonstrated in Figure 2, this coincides with the region that the infotaxis agent navigates towards in the absence of any detected hits.

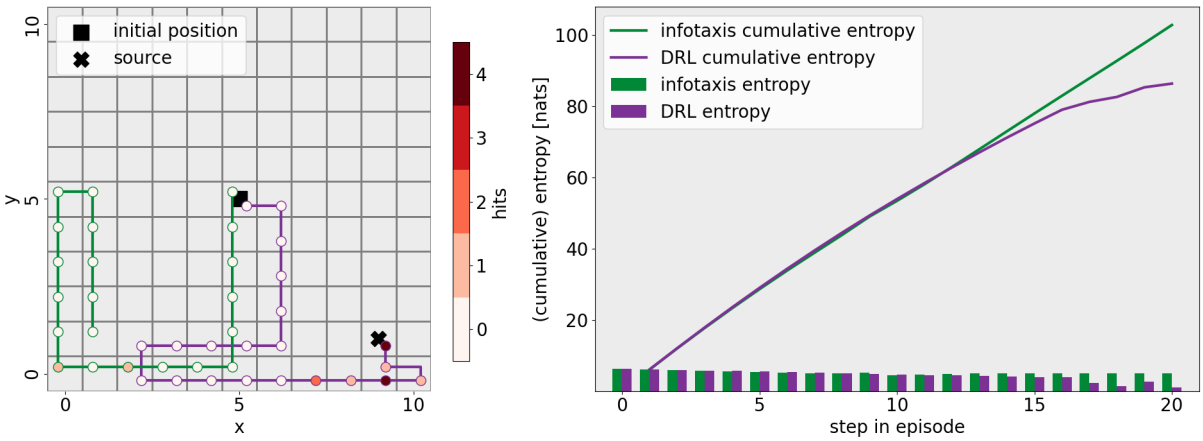

Figure 2. Agent navigation in an environment with $\tilde{V} = 2$ and $\tilde{D} = 2$, and source term parameters $\tilde{\mathbf{x}}_s = (9, 1)$ and $\tilde{\phi} = 2$: (left) DRL and infotaxis sampling paths, and (right) (cumulative) entropy. Lower entropy indicates a reduced uncertainty in the estimation of the source term parameters.

Table 1. Success rate of DRL and infotaxis evaluated across different flux values $\tilde{\phi}$ (training encompassed all source term scenarios).

	$\tilde{\phi} = 1$	$\tilde{\phi} = 2$	$\tilde{\phi} = 3$	$\tilde{\phi} = 4$	$\tilde{\phi} = 5$
DRL	40%	65%	80%	84%	85%
infotaxis	31%	49%	65%	72%	76%

4.3 Sensitivity to $\tilde{V}$ and $\tilde{D}$

Figure 3 illustrates a relative Discrete Ranked Probability Score [19] ($\text{DRPS}_{\text{posterior}}$ / $\text{DRPS}_{\text{prior}}$) for DRL and infotaxis policies across environments with varying known diffusivities $\tilde{D}$ and wind speeds $\tilde{V}$. The agents are trained under these specific environmental conditions. DRPS is a measure for belief state accuracy, favoring lower scores. Note that $\text{DRPS}_{\text{prior}}$ is the initial uniform belief state and remains consistent across all test cases. We observe improved estimations of source location in conditions with higher winds. In cases where $\tilde{V} > 0$, DRL outperforms infotaxis. Additionally, we observe that the estimation of source location on the crosswind axis ($\tilde{x}_s$) surpasses that on the along windward axis ($\tilde{y}_s$), and that lower diffusivities are beneficial for STE in such cases. Under isotropic conditions ($\tilde{V} = 0$), DRL and infotaxis perform comparably given the specified training parameters.

5 Discussion

The findings presented in this study are a first effort in applying DRL with an information-based reward to solve the STE problem with mobile gas sensors. While infotaxis adopts a myoptic perspective, maximizing information gain solely for the upcoming action, DRL has the capability to develop a far-sighted strategy for action selection, considering information gain across the entire sampling trajectory. This study demonstrates that adopting a DRL approach has the potential to improve the accuracy of STE.

Our findings show that DRL outperforms infotaxis in non-isotropic plume environments. The policy allows the agent to select actions that might not be immediately optimal, but are chosen in anticipation of achieving better rewards in the future. Comparative analysis shows that the DRL agent opts to stay (downwind) centralized when its belief over the source term parameters remains uncertain. In contrast, the infotaxis agent tends to move towards the edge of the domain under similar circumstances. This trajectory could bring the infotaxis agent closer to the source and valuable information, but also poses a risk of leading the agent in the opposite direction, further away from the source and information about the source term parameters.

DRL's superiority over infotaxis does not extend to isotropic plumes in this study, possibly due to the increased symmetry of the hit map and hence limited directional information. In contrast, recent work by Loisy and Eloy [9] showed DRL's efficiency in source localisation within isotropic plumes. Key distinctions between our studies are that [9] exclusively focuses on localisation, omitting flux estimation. In their case, the flux strength was known which assisted the agent in source seeking, whereas in our study the search for informative observations is further complicated by the unknown source strength. Their policy employed a -1 reward, terminating episodes upon the agent's arrival at the source - information not available to our agent. Moreover, episodes initiated with the agent receiving a hit, while episodes can start with zero hit observations

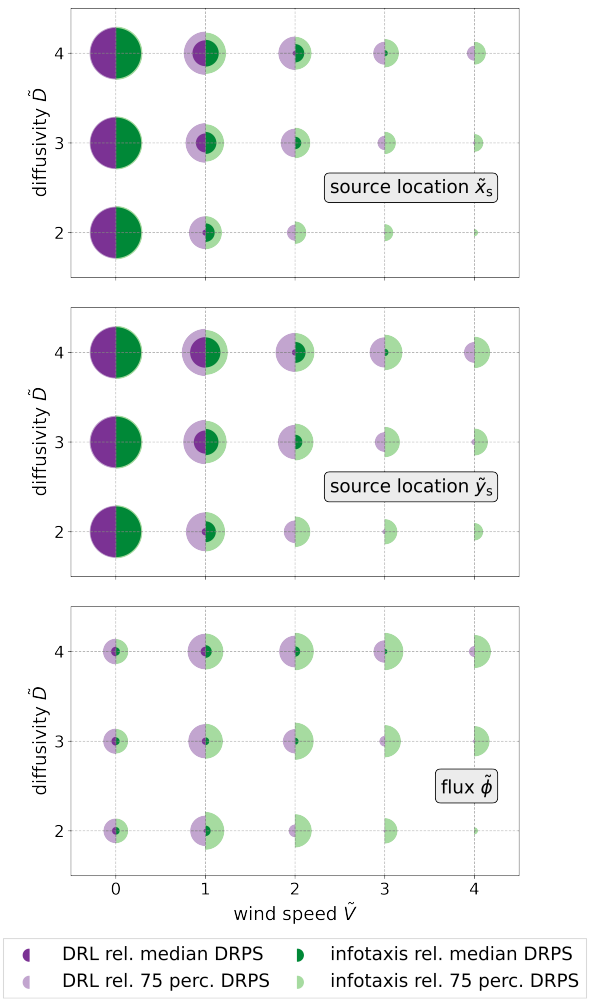

Figure 3. Relative Discrete Ranked Probability Score (DRPS) for source location $\tilde{x}_s$ (top), $\tilde{y}_s$ (middle) and flux $\tilde{\phi}$ (bottom). Circle radii correspond to the median (dark shade) and 75th percentile (light shade) relative DRPS. DRL results are shown as (purple) left half circles and infotaxis results are shown as (green) right half circles.

in our study.

Loisy and Eloy [9] find that the performance of space-aware infotaxis is close to the optimal policy obtained by DRL for source seeking. This variation on traditional infotaxis incorporates an expected distance measure between agent and source, and is designed to efficiently select the shortest path to the source. Our approach differs: we aim to strategically select a 20-step path that optimally constrains the source term parameters through observations. In the environment studied currently, the optimal policy indeed leads the agent towards the source. Nonetheless, unlike Loisy and Eloy [9], our agent entertains multiple possible flux strengths, such that the agent may place equal probability to the source being close by with a low flux strength as it being further away with a higher flux strength. Furthermore, we deliberately refrain in this study from rewarding actions that select the shortest path upfront, as our long-term aim is to develop a framework that can perform in various environments. In particular, future work will include domains with multiple sources. In this case, we anticipate that the shortest path to one of the sources is not the optimal policy to constrain the uncertainty over the parameters characterising all sources.

6 Outlook

As we look ahead to the practical implementation of our DRL framework, we identify several challenges and possibilities: (a) The framework presented in this study is trained for known wind conditions and diffusion. Real-world application requires training across a range of atmospheric conditions, which can be inferred from observations by mobile and ground-based sensors, or a combination of the two [20]. (b) The current observation model (1) assumes steady-state atmospheric conditions, a simplification that may not hold in real-world scenarios. More complex, time-dependent models, as employed by Pirk et al. [20], may enhance the accuracy of the inference results. However, these models are computationally expensive, posing challenges for online usage. (c) We currently assume the presence of a single source within the domain. However, in reality, each "grid cell" represents a possible greenhouse gas source or sink. Future research aims to extend the framework to map heterogeneous greenhouse gas fluxes across a domain. (d) Applying the framework in field studies requires upscaling to larger domains and parameter spaces. Future work should investigate whether the FC and CNN architectures in this study can sustain their comparable performance within a realistic computational budget.

Despite the anticipated challenges in implementing the framework in real-world scenarios, our study represents a significant advancement in the methodology for STE of gas sources. It demonstrates encouraging outcomes achieved through the guidance of drones by information gain strategies.

Acknowledgements

This work was supported by the Research Council of Norway (project 301552 "Upscaling hotspots – understanding the variability of critical land-atmosphere fluxes to strengthen climate models (Spot-On)" and project 333232 "Strategies for Circular Agriculture to reduce GHG emissions within and between farming systems across an agro-ecological gradient (CircAgric-GHG)"). This work is a contribution to the strategic research initiative LATICE (Faculty of Mathematics and Natural Sciences, Uni-

versity of Oslo, project UiO/GEO103920) as well as the Centre for Computational and Data Science (dScience, University of Oslo).

References

[1] E. A. G. Schuur, A. D. McGuire, C. Schädel, G. Grosse, J. W. Harden, D. J. Hayes, G. Hugelius, C. D. Koven, P. Kuhry, D. M. Lawrence, S. M. Natali, D. Olefeldt, V. E. Romanovsky, K. Schaefer, M. R. Turetsky, C. C. Treat, and J. E. Vonk. "Climate change and the permafrost carbon feedback". In: *Nature* (2015). DOI: 10.1038/nature14338.

[2] M. Hutchinson, H. Oh, and W.-H. Chen. "A review of source term estimation methods for atmospheric dispersion events using static or mobile sensors". In: *Information Fusion* (2017). DOI: 10.1016/j.inffus.2016.11.010.

[3] S. Särkkä and L. Svensson. *Bayesian Filtering and Smoothing*. 2nd ed. Cambridge University Press, 2023. DOI: 10.1017/9781108917407.

[4] A. Francis, S. Li, C. Griffiths, and J. Sienz. "Gas source localization and mapping with mobile robots: A review". In: *Journal of Field Robotics* (2022). DOI: 10.1002/rob.22109.

[5] M. Vergassola, E. Villermaux, and B. I. Shraiman. "'Infotaxis' as a strategy for searching without gradients". In: *Nature* (2007). DOI: 10.1038/nature05464.

[6] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, 2018.

[7] B. P. Duisterhof, S. Krishnan, J. J. Cruz, C. R. Banbury, W. Fu, A. Faust, G. C. H. E. de Croon, and V. J. Reddi. *Learning to Seek: Autonomous Source Seeking with Deep Reinforcement Learning Onboard a Nano Drone Microcontroller*. 2019.

[8] L. Wang, S. Pang, and J. Li. "Olfactory-Based Navigation via Model-Based Reinforcement Learning and Fuzzy Inference Methods". In: *IEEE Transactions on Fuzzy Systems* (2021). DOI: 10.1109/TFUZZ.2020.3011741.

[9] A. Loisy and C. Eloy. "Searching for a source without gradients: how good is infotaxis and how to beat it". In: *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* (2022). DOI: 10.1098/rspa.2022.0118.

[10] Y. Zhao, B. Chen, X. Wang, Z. Zhu, Y. Wang, G. Cheng, R. Wang, R. Wang, M. He, and Y. Liu. "A deep reinforcement learning based searching method for source localization". In: *Information Sciences* (2022). DOI: 10.1016/j.ins.2021.12.041.

[11] A. Loisy and R. A. Heinonen. "Deep reinforcement learning for the olfactory search POMDP: a quantitative benchmark". In: *The European Physical Journal E* (2023). DOI: 10.1140/epje/s10189-023-00277-8.

[12] M. Park, P. Ladosz, and H. Oh. "Source Term Estimation Using Deep Reinforcement Learning With Gaussian Mixture Model Feature Extraction for Mobile Sensors". In: *IEEE Robotics and Automation Letters* (2022). DOI: 10.1109/LRA.2022.3184787.

[13] A. van Hove, K. Aalstad, and N. Pirk. "Using reinforcement learning to improve drone-based inference of greenhouse gas fluxes". In: *Nordic Machine Intelligence* (2023). DOI: 10.5617/nmi.9897.

[14] L. P. Kaelbling, M. L. Littman, and A. R. Cassandra. "Planning and acting in partially observable stochastic domains". In: *Artificial Intelligence* (1998). DOI: 10.1016/S0004-3702(98)00023-X.

[15] D. J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.

[16] M. J. Kochenderfer, T. A. Wheeler, and K. H. Wray. *Algorithms for decision making*. The MIT Press, 2022.

[17] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis. "Human-level control through deep reinforcement learning". In: *Nature* (2015). DOI: 10.1038/nature14236.

[18] A. Loisy and C. Eloy. "OTTO: A Python package to simulate, solve and visualize the source-tracking POMDP". In: *Journal of Open Source Software* (2022). DOI: 10.21105/joss.04266.

[19] G. Candille and O. Talagrand. "Evaluation of probabilistic prediction systems for a scalar variable". In: *Quarterly Journal of the Royal Meteorological Society* (2005). DOI: 10.1256/qj.04.71.

[20] N. Pirk, K. Aalstad, S. Westermann, A. Vatne, A. van Hove, L. M. Tallaksen, M. Cassiani, and G. Katul. "Inferring surface energy fluxes using drone data assimilation in large eddy simulations". In: *Atmospheric Measurement Techniques* (2022). DOI: 10.5194/amt-15-7293-2022.