Using Model-Free Reinforcement Learning Combined With Underwater Mass Spectrometer and Material Archiving Coupled to Lab Analysis for Autonomous Chemical Source Verifications

Connor Tate
Florida Institute for Human and Machine Cognition
University of West Florida
Pensacola, USA
ctate@ihmc.org

David P. Fries
Florida Institute for Human and Machine Cognition
University of West Florida
Pensacola, USA
dfries@ihmc.org

Micael Vignati
Florida Institute for Human and Machine Cognition
University of West Florida
Pensacola, USA
mvignati@ihmc.org

Kevin Francis
University of West Florida
Somerville, USA
kjf13@students.uwf.edu

***Abstract*—To support the general problem of Autonomous Underwater/Surface Vehicle (AUV/ASV) based chemical detection, source localization, we propose the design of a system that is a fusion of AUV/ASV with Q-learning, and a real-time underwater mass spectrometer, used to provide the feedback and reward signal for in situ source localization. Additionally, an autonomous sampler can be coupled to the system permitting molecular material archiving for subsequent expanded measurement and validation in the lab. This real-time chemical sensor and archived sample capture and verification approach yields an adaptive sensing and sampling system. The in situ mass spectrometer allows for real time measurement of membrane compatible chemistries such as volatile oxidative compounds (VOC's) and lightweight gases, while the sampler purifies, enriches and accurately isolates targeted molecular compounds in the field for subsequent full mass spectrometer analysis back in the lab. In the overall AUV system design, the battery driven mass spectrometer provides real-time mass spectrometer signals for reinforcement learning (RL) behaviors and the portable adaptive sampling system automates sample collection, molecular purification/concentration and preservation. The mass spectrometer is of the membrane inlet type and the automated sampler system is a combination of customizable fluidic management systems, pumps, valve arrays and motion control systems. For the field sampling use, the prototype sampling module is designed for triggered sensing and sampling but also can be variably actuated to sample variable volumes over any period of time. The mass spectrometer and sampling systems can be hosted on AUVs/ASVs for most chemical source localization activities. The entire mobile system: AUV mobile platform, reinforcement learning controller, mass spectrometer, and sampler, constitute an adaptive chemical sampling platform. The 'back end' laboratory identification is performed using any type of mass spectrometers and can provide a high confidence verification of the specific material archived. The results from the lab verification can also constitute the design of a reward signal for subsequent Q-learning training, mass spectrometer data sub-system to increase the accuracy of the source localization policy. The potential of using mass spectrometer data to train a Q-learning based agent allows the team to pretrain the agent with real sensory data similar to that which will be seen in the field for future deployments. Appropriately simulated data can approximate the environment and distribution patterns that are anticipated for the development of a custom reward function, representative of the mission objective. Preliminary simulations testing the agent's performance, utilizing a trained policy in a similar environment in which the location of a generic 'pollution source' has been perturbed from the training scenario, have shown promising results. The policy is acquired by training on pollution data for a set environment in which the trade-off between exploration and exploitation is defined appropriately for the environment size, pollution distribution and training duration to optimize the agent's learning. That policy is then tested in a similar but slightly perturbed environment. This method can be applied to future missions to allow for continual policy update based on the observed data. This would be an advantageous approach as it limits the necessity for operator-vehicle communication giving the agent sufficient autonomy to locate the source based on its prior training as well as circumvents the need for a model-based decision and control approach as the agent becomes better trained through real world observations. This is a model-free learning approach requiring no a priori knowledge of the environment. This has a distinct benefit over model-based approaches which are dependent on the accuracy and fidelity of the environmental model during the training of the agent, which is notoriously difficult both logistically and computationally.**

***Index Terms*—Underwater mass spectrometry, Q-learning, Source Localization, Reinforcement Learning**

I. Introduction

Source localization is a technical challenge spanning multiple domains: environmental monitoring of pollution, identification and assessment of hydrocarbon sources and leaks for

global climate budgeting, as well as commercial and military applications. The capability of identifying a chemical source and analyzing the chemical signature for fingerprinting can aid in the assignment of responsibility for contamination or the assessment of a chemical's ecological and biochemical impact. Historically, AUV based methods for detecting, localizing, and assessing areas of interest in the marine environment rely on lawnmower, yo-yo or spiral patterns in order to cover a broad space within the region of interest. These brute force methods are effective for mapping a space, however they are inefficient in their use of time and manpower. When detecting and analyzing a signal is time sensitive, such as in the emission of unwanted pollutants, the ability to set out with a mission and adapt to in situ conditions is essential. Oftentimes AUV's are equipped with sensor suites that allow for high resolution data throughput and post analysis to tackle an array of monitoring mission objectives. Many of these sensors such as underwater mass spectrometers (UMS), dissolved oxygen (DO), fluorescence, backscatter and other optical sensors can be directly linked into the vehicle control board to provide rich sensory information. By utilizing these sensor suites, the AUV gains increased capacity to perceive its environment and make decisions in a framework that is related to the mission objective. By implementing reinforcement learning, an autonomous vehicle can become a more intelligent agent capable of sensing relevant environmental parameters and learning from them through training and rewarding optimal outcomes. Training can occur prior to mission deployment and be informed by the literature or a subject matter expert in order for the agent to develop an optimal policy given a certain environment and mission objective. To support AUV based chemical detection and source localization, Q-learning and a real-time underwater mass spectrometer are combined to provide the feedback and reward signal for in situ source localizations. Real-time input from onboard sensing will allow the agent to update its policy based on the true in situ observations. The observed UMS data combined with physical samples can be further analyzed in the lab providing ground truth for improved training and future mission deployments. Initially the training environment can be developed sufficiently simplistic, capturing the relevant parameter relationships and patterns that will inform achievement of a goal while remaining independent of the training environment model. This generalized simplicity aims at providing better policy transferability to real world testing while avoiding over-fitting. Using a hybrid Q-learning plus observed data approach the agent is trained to recognize key patterns in incoming data from the UMS and other on-board sensors and integrate these true values into the policy for future deployment, allowing for better real world performance. The coupling of observed sensory data with physical sample acquisition can be used to verify chemical signatures for deeper analysis providing insight into unique ecosystem impacts as well as future research directives. To demonstrate the process we simulate a mission in which the agent must locate the source of a methane ebullition modeled after domain knowledge based on literature review and evidence based on real data acquired from United States Geological Survey (USGS). Prior to testing we will train the agent using a simplified and generic environment with methane and oxygen profiles modeled after insights gained from literature. The agent is trained for a number of epochs satisfactory to develop an optimal methane source localization policy, and then tested in the simulated environment.

II. Background

Traditionally, the problem of source localization has been tackled from stationary sensing nodes, physical sample collection by field teams, towed sensor arrays behind ships or more recently with the use of autonomous mobile platforms. Autonomous vehicles provide more flexibility than other localization modalities due to their ability to reach remote and extreme environments with reduced manpower, the array of equipped sensors as well as endurance. Whereas other methods for detecting and localizing require extensive manpower, resources and time succumbing to spatial and temporal limitations. The variety of vehicles available for autonomous monitoring and sensing range from gliders, surface vehicles, lagrangian drifters and torpedo shaped submersibles; this array of platforms enables multiple approaches toward solving the source localization problem. However current deployment of these vehicles is constrained by predefined trajectories and limitations in operator-vehicle communications that prevent trajectory modification [1]. In the marine environment communication is bandwidth limited and in some instances such as localization for mine counter measures, communication and human intervention for trajectory modification is restricted therefore a need for increased autonomy and higher levels of system intelligence are essential [2]. The importance of expanding the functionality and applicability of these vehicles in low communication environments with varying levels of uncertainty has been identified as a key research area.

A. Source Localization Approaches

In the past 20 years research in the area of autonomous source localization and vehicle control by means of chemical profiling via on-board sensors has increased significantly. Evidence supporting on-board chemical profiling for plume tracing and source localization can be found in many field studies [3], [4], [5], [6], [7]. However, of the field demonstrations, most are limited to the traditional lawnmower and yo-yo patterns designed to cover a large area. Although effective for mapping large swaths these methods of trajectory planning are inefficient when considering the problem of source localization [1]. Recent work integrating on-board sampling for adaptive trajectory control focuses on one of three areas: targeted features of interest (TFOI), objectives of sampling mission (OSM), and multi-vehicle networking [6]. Some related efforts involve adaptive sampling techniques which include informative path planning strategies such as those employed by [8], [9] integrating sensory information with the vehicles working environment model to update the trajectory. One such approach is that of [1] in which they use

real-time information-seeking algorithms to optimize source localization tasks by having the agent autonomously decide sampling locations, demonstrating promising test results in the field.

Outside of limited field demonstrations, much of the current research is simulated with the majority providing simulations in a two dimensional grid [6], [10], [11], [12]. This limitation is typically due to computational constraints and the increase in complexity and computational demands of simulating an environment in 3D.

In the works reviewed, methods for source localization include: supervised learning methods [13], chemotaxis and Fisher Information Matrix [10], and a range of RL approaches [14], [11], [6], [12]. Of the RL methods reviewed three contributed to vehicle control and spatial navigation related to matters of depth control, obstacle avoidance and propulsion. As it is related to this work, the papers by Weidemann and Wang [6], [12] implemented and supported the use of RL for the subject of source localization. As it stands, there are limited applications of RL in the domain of chemical source localization and of those that do employ it many provide only 2D simulations. 3D simulation or field work usually relies on model-based learning which is subject to model fidelity sensitivity during training, which is notoriously difficult due to the complexity and unpredictable dynamic nature of the marine environment. The additional constraints introduced by vehicle communications and the proven need for increased autonomy means that online mission reconfiguration is limited [2]. The method presented in this paper aims at overcoming the shortcomings of traditional and autonomous source localization methods such as chemotaxis and anemotaxis by training on multiple environmentally relevant parameters in a 3D simulated environment informed by the expectation of a subject matter expert.

B. Chemo Sensing and Sampling (mass spectrometer and Autonomous Sampling Systems)

Chemical signals in the field are unlike light and sound signals. Chemical information is multidimensional; that is, each chemical channel (i.e., parent molecule) contains added chemical information available in the fragments that make up the molecule. This applies in both atmospheric and underwater chemical dynamics. Ideally, a sensor suitable for such numerous input signals would be capable of flexible, simultaneous signal detection with unique chemical identification information. The Mass spectrometer is arguably [3] the most versatile of chemical sensors and fulfills these requirements. In situ mass spectrometry system is capable of real-time, dynamic, in-water field analysis. The platform shown in "Fig. 1" typically operates at 70 watts and has been deployed on larger, limited endurance AUV's, however lower power versions are needed to permit placement on the new class of renewable power based AUV's to create large spatio-temporal maps which currently do not exist. Mass spectral data are multispectral data and are commonly visualized as standardized x-y graphical imagery but can be converted to 2D array data. Traditionally,

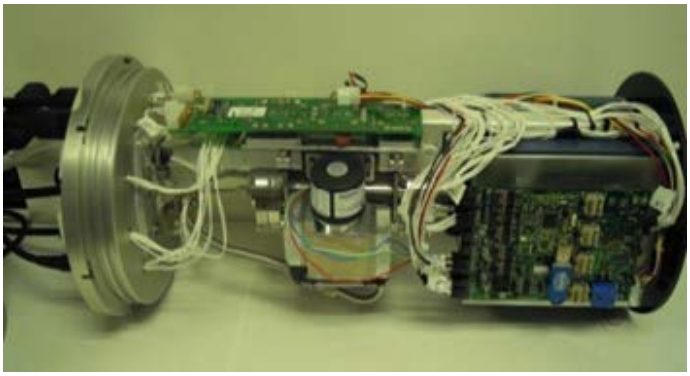

Fig. 1. Membrane Introduction mass spectrometer, Power: 60-80 Watts, Voltage: 24 VDC, Dimensions: Length: 64 cm Diameter: 24 cm Weight: In air: 35 kg, in water: 5 kg negatively buoyant, Depth rating: 2000 m.

data is processed through spreadsheets and csv file formats but processing algorithms for time, (mass-charge), and intensity can also be used as a signal processing approach. In most of these approaches, time and power consuming numerical calculations are carried out for mass spectrometer peak ratios, differentials, and background subtractions. Referencing and searching of known peak patterns for chemicals (fragment patterns) are also applied to the resultant spectral data, which can help to discern the interfering "picket fence" of simultaneous mass peaks detection. Usually in standard mass spectrometer systems a pre-separation step is applied (e.g. chromatography) to separate the incoming chemicals in time and space prior to mass spectrometer detection. This added analytical module is extremely power prohibitive and is rarely seen in field-able systems, especially with in situ UMS aquatic measurements. An alternative method for low power AUVs would be to find higher order relationships between spectral signals for AUV adaptive sampling control. With environmental real-time UMS data, a highly desired goal is generating potential indicators that might help in the classification of a specific state of the environment, (e.g. the suitability for hosting biological organisms). Additionally, there is interest in finding the distance between sample and reference spectra, or for distances between the early and late stages of acquired mass spectral signals. Using Q-Learning may allow low power, real-time, embedded global optimization processing for vehicle control and is an improvement over model dependent remote processing schema [15] and cloud-based approaches [16] which can be laborious and time consuming. In addition to the case example of chemical sensing and autonomous system control, the proposed method can be extended into other AUV applications and other sensor systems; e.g. pH sensing for mapping ocean acidification, underwater object recognition, intelligent navigation and control, chemoperception modeling [17] and can be applied to the larger field of mass spectrometry, in general, for new signal processing and detection schema. Current technology readiness level of mass spectrometer technology permits the real-time detection of certain classes of chemical compounds (permanent gases, volatile,

semi-volatiles) which limits the characterizations possible of the underwater chemical space for environmental, industrial or ecological chemical dynamic studies. A supplemental approach is to combine mass spectrometer data with automated physical sampling of targeted classes of chemicals (e.g. non-volatiles, elemental, large organics, ecological chemo signals) for capture and archiving for post processing shore-side or ship side to allow for these unattainable chemistries to be mapped and understood. By utilizing automated in situ chemo-selective sampling capable of both time-series sampling and accurate sequestering of material fluxes and transformations across diverse spatial scales, a great deal of traditionally lab-based effort can be applied in the field directly at the source of sampling. This point-source analyte preparation also leads to improved sample robustness and data reliability. The technology suite described here allows for fast sampling and slower, high content analysis leading to more information about environmental change. This system in effect permits an extension of the radius of lab based mass spectrometers without these unwieldy units having to be made transportable or miniaturized. To supplement the real-time mass spectrometer a automated fluidic processing system has been developed which allows for selective purification, extraction, and archiving of chemicals within an AUV host, providing captured samples to be transported back to a lab-based mass spectrometer for detection and confirmation "Fig. 2". Initially, the mobile sampling platform development and verification has been directed toward in situ hydrocarbons and associated oil contaminants. The system also contains an embedded fluorometer that can be configured for oil sensing and triggering of the sampler and archiving of materials. The entire mobile system constitutes an adaptive sampling platform. The 'back end' laboratory identification is performed using typical mass spectrometry and other analytical techniques that can be applied to the archived material, allowing for optional interdisciplinary study and confirmations of sampled archived analytes. The mass spectrometer and sampler are modular designs and are interfaced and linked with an embedded controller for signal communication and control to allow trigger signals from the real time mass spectrometer (or optional embedded fluorimeter) to initiate an automated sampling and archiving cycle. The command signal is a variable output depending on learned or mission defined mass spectrometer sensing criteria which allows for adaptive sensing and sampling. Various sense and sampling control schema can be configured (e.g. centralized controller; distributed communication/commands; external sensors - such as CTD), that are correlated with the mass spectrometer real-time data.

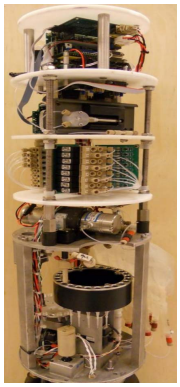

Fig. 2. The robotic sampling system, including fluidics manager, sample concentrator and embedded fluorescence detector for oil sensing/screening (optional). Hardened underwater housing removed for illustrative purposes.

III. Methods

A. Mass Spectrometry and Fluidic Sampling

In the AUV system design, the battery driven mass spectrometer provides real-time mass spectral signals for determination of policy driven behavior. The mass spectrometer detects specific hydrocarbon concentrations with the capacity for detecting multiple compounds for many different applications. The 'front end' sampler-purifier is based upon fluidic processor technology and automates sample collection, filtration, fixing, labeling and chemical purification and concentration by utilizing a combination of customizable and battery-driven, fluidic management systems, pumps, valve arrays and motion control systems. Currently, the prototype module is designed for up to seven-day deployments, where it can collect variable sample volumes and perform 25 sequential sample treatments. The system also contains a small format embedded fluorometer that can be configured for oil sensing and adaptive triggering of the sampler and archiving of materials. The sampler is a fully functioning system that is selectively open to the environment in which it is deployed for automatic mass transfer of environmental material. The sampler system is additionally open to tuning and reconfiguration for detection of different materials through the design and engineering of specific recognition/capture probes that are an integral part of the extraction subsystem. In marine waters and other harsh environments, rapid and sequestered sample preparation and extraction is vital in the reduction of noise interference and signal enhancement. A prototypical sampling configuration for hydrocarbon analysis (e.g. during oil-spill events) includes sample columns containing custom developed fluorinated C18 functional capillaries. Upon retrieval, columns are analyzed via direct injection in a typical gas chromatography-mass spectrometer (GC/MS). To evaluate the efficiency of columns to capture hydrocarbons, standards were prepared with dilute carbazole in water. Carbazole was spiked in local seawater from Bayboro Harbor to simulate in-situ conditions [22]. Additional experiments were performed using an underwater membrane introduction mass spectrometer (MIMS) for analysis of b-cyclocitral, an indicator molecule of the freshwater harmful algal bloom (HAB) forming organism Microcystis aeruginosa, which secretes the hepatotoxin microcystin in freshwaters, posing health problems regulated by water resource managers and health department officials in numerous jurisdictions. Other samples collected using the AUS system, including brevetoxin (saltwater HAB toxin) , chlorophyll (typical blue-green algae indicator) and chemical pollutants

such as bisphenol and PCBs were analyzed by others using typical mass spectrometry techniques. For the purposes of this demonstration the mass spectrometer is assumed to detect a single generic methane compound typical of methane seeps. Due to the complexity of methane dissolution, its effect on the surrounding environment and subsequent dissolved gas concentrations, we simulated a toy training environment mimicking a simplified generic distribution pattern informed by the literature available for the area around the selected test environment. In this demonstration the agent learns multiparameter gradient patterns in relation to methane and dissolved oxygen for the localization of a methane seep. The test environment is based on the USGS data from a May 2016 dive utilizing a Bluefin-12 AUV to map, analyze and locate a passive margin methane seep located off the coast of Miami on the continental slope above the hydrate stability zone [18].

B. Simulation

The training and testing environment were generated in python 3.9 using numpy and matplotlib libraries for indexing the environment values and visualizing the data.

a) Training Environment: For the training environment "Fig. 3", the methane distribution shape is modeled based on a gravity driven dispersion that is a predominant governing factor for biogenic passive margin seeps above the hydrate stability zone [18]. The complexity of the chemical distribution and size of the training environment are scalable and dependent on the modelling and computational capacity at time of training. Due to the complex and dynamic nature of methane diffusion and gas solubility laws a generalized logarithmic curve was applied to simulate rapid consumption of methane near the seafloor with lower concentrations occurring at the boundaries and surface. The dispersion gradient is calculated in relation to the location of the source. Methane is present in its highest concentration at the source and then follows a logarithmic diffusion from source to surface. An increase in surface concentration of methane is informed by [19] in which increased levels of background methane were seen in the mixed layer.

The dissolved oxygen diffusion gradient is determined by a model fit to the vertical oxygen profile observed in the region. The vertical profile was established using a subset of the GEOMECC-2 cruise data which includes transects for multiple depths near to the test environment. These data files were averaged to produce an average vertical profile for oxygen in a region with similar topological and current conditions to the test environment. From this average a 5th order polynomial regression was performed to develop the oxygen diffusion model for the training environment "Fig. 4". The resulting model is:

$$y = 38.333x^5 - 112.65x^4 + 121.33x^3 - 55.798x^2 + 8.2672x + 0.6007 \quad (1)$$

b) The Agent and Q-States: The agent is trained to detect the gradients observed in the methane data received from the mass spectrometer, as well as the dissolved oxygen for each sampling point. These data inputs are normalized to scale to different environments. From this the agent learns the relationship between dissolved oxygen and methane in order to choose the best action that will lead it to the location of the seep and avoid mistaking local minima and maxima. The agent chooses between six available actions up, down, left, right, forward and backward. During training the action chosen is determined by the Epsilon-Greedy Algorithm, either random or selected based on maximum perceived reward for each possible state. In this demonstration a state is composed of the methane gradient and oxygen gradient. The goal state is the one with the maximum reward calculated using the Bellman Equation which is the foundation for this Q-Learning approach. This algorithm allows the agent to estimate the possible future reward obtained from an action resulting in a new state. This (state, action) pair is stored with the associated reward value in the agent's Q-table "Fig. 5". As the agent moves through the environment the agent updates the value for the discrete pairs based on observed actual reward. This eventually converges in the optimal policy.

For this agent both positive and negative reinforcement

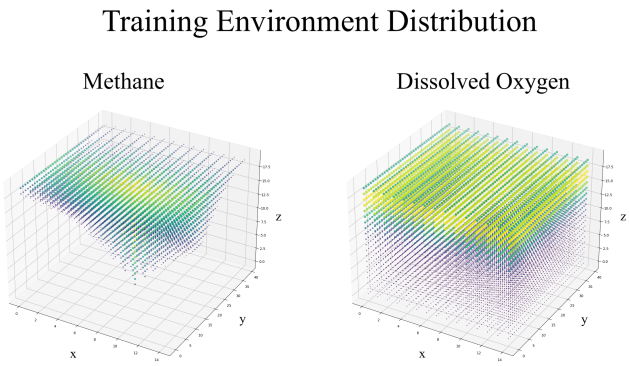

Fig. 3. Training environment showing the methane distribution from source with a logarithmic decline reaching the surface **left** and the dissolved oxygen distribution following the regional vertical oxygen profile **right**.

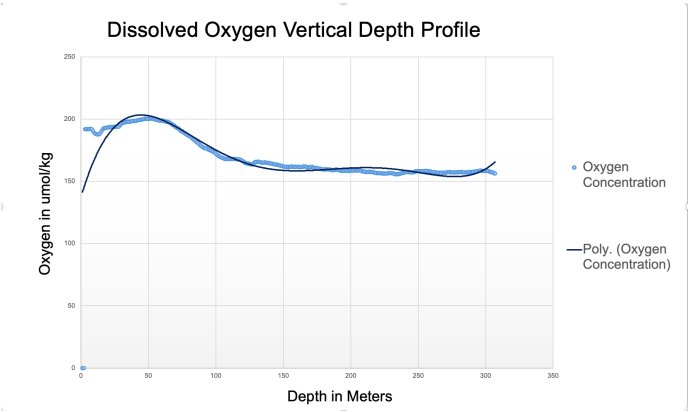

Fig. 4. The defined vertical oxygen profile developed for the training environment from the GEOMECC-2 cruise data.

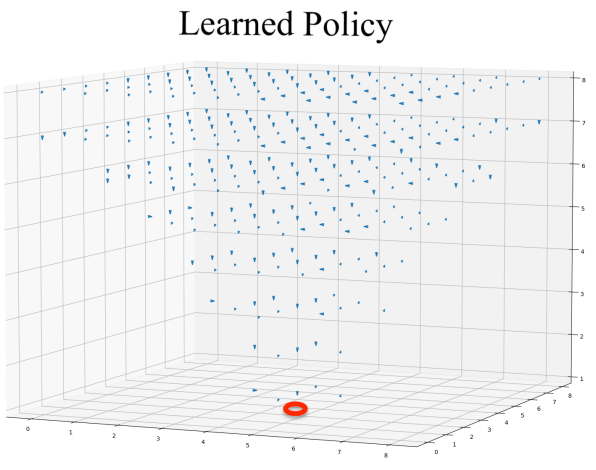

Fig. 5. Demonstrates a smaller version of the Q-table for clarity. The image is composed of arrows pointing in the direction of the best move determined by the methane and dissolved oxygen gradients experienced at that location, with all moves converging to the location of the source.

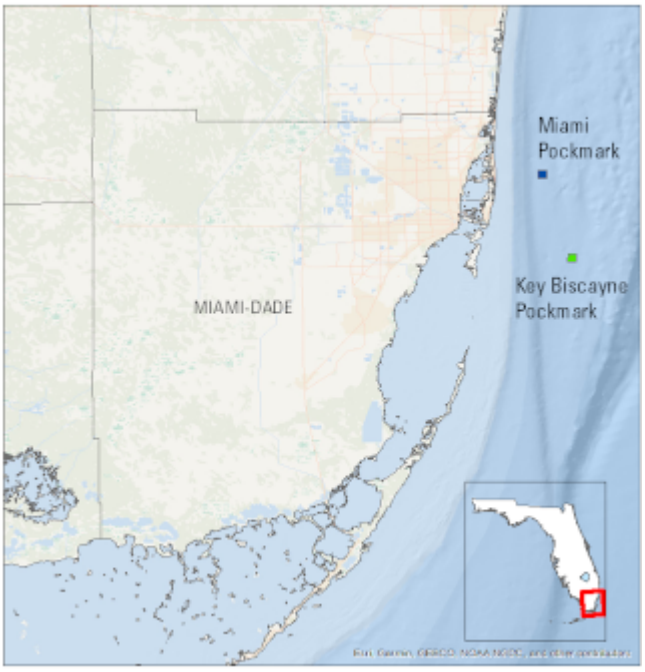

Fig. 6. Image Courtesy of Cunningham and Wescott, showing sampling location of the passive margin shallow seep location at Miami Biscayne Pockmark.

are implemented to encourage progression towards the goal state, location meeting the methane concentration maximum. Positive reward is only achieved at the location of the source, while accumulating negative reward occurs for each step taken that does not result in the goal state. The implementation of a negative rewards encourages the agent to find the most direct path and not rely on the the first paths found. For this agent the size of the training environment mimics the size of the test environment and is scaled to the sampling frequency exhibited in the USGS mass spectrometer data. This data was acquired using a UMS with a 5 second sampling frequency, attached to a Bluefin-12 AUV with a survey speed of 6 Knots. The resulting environment has dimensions 15 x 40 x 20, this results in 12,000 possible sampling locations. A state has 6 parameters, three for the methane gradient in each axis (x, y and z) and three for the oxygen gradient in each axis; the gradients are each discretized into three buckets (positive, negative and neutral) for estimating Q-states, resulting in 216 discrete states.

IV. Results

A. Testing

The testing environments were inspired by dive data from Miami Pockmark and Key Biscayne Pockmark, “Fig. 6” [20]. The dataset provides x, y for specific latitude, longitude and depth with methane values in mmol/kg, and dissolved oxygen in mmol/kg.

Each location consisted of two separate dives in which the vehicle mapped the seep location at an approximate depth between 290-305 meters below sea level. Those two dives for each location were combined to create one dataset for each location. This was then visualized to determine directionality and shape of the distribution using least squares regression. The resulting output was for the sampling depth attained during the 2016 dive. This demonstrated a fairly linear distribution over narrow depth provided. We supplemented the diffusion in relation to depth of these parameters using a perturbation of our training diffusion pattern. As can be seen in “Fig. 7”, the training environment for methane distribution presents an even diffusion throughout the space, this ensures the agent experiences all state spaces with 360 degree access to the seep in the survey area. In the testing environment the location

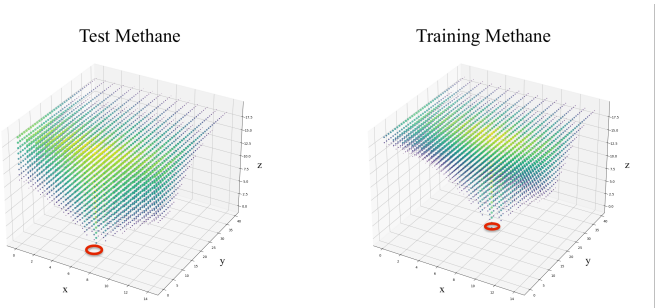

Fig. 7. **Right** Training environment showing methane distribution, dissolved oxygen distribution retains the distribution previously discussed in prior sections. **Left** Testing environment showing the methane plume existing partially with in bounds with a different shape and diffusion that that of the training environment.

and distribution of the methane is perturbed by changing its position within the survey field and limiting its diffusion within bounds to mimic the distribution seen in the USGS data, in which the survey area bounds cover the seep location but slice

through the plume which has a shore-ward flow due the the up-welling occurring from the proximity to the Gulf Stream and location on the continental slope. For this scenario the agent performed 10,000 trials per training session, ten training sessions were performed resulting in ten distinct policies. Each trial consisted of 500 moves in which the agent could explore the environment and update its policy. From the resulting policies the best, "Fig. 8" was selected for deployment in the testing environment. At the initiation of the trial the learning parameters were set as follows:

- Learning rate = 0.1
- Epsilon = 1
- Epsilon Decay = 0.999
- Reward = 100
- Negative Reward = -1 / step
- Discount Factor = 0.9

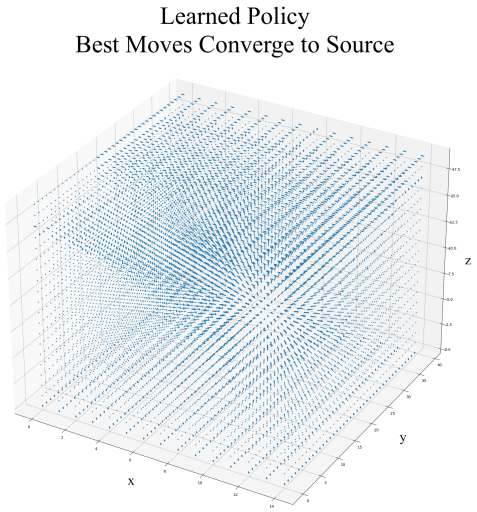

Fig. 8. The best developed policy after 10 training sessions consisting of 10,000 trials. This figure depicts the direction representing the best move based on the methane and oxygen gradients experienced in each location.

Once the epsilon had become tangential to zero the agent only made on policy decisions "Fig. 9". At the end of training the best policy exhibited maximum reward at the 6000 trial mark, with the following trials resulting in maximum reward each run. Attaining the maximum reward, the adjusted reward based on the agent's relative starting location compared to the goal state, is indicative of the agent learning the most direct path to the location of the source.

Once the best policy had been attained we deployed the agent in the testing environment. For testing purposes we assumed a sea surface deployment in which the agent started within background levels of methane. The number of steps the agent could take within a survey was constrained to 1000, approximating the steps taken by the Bluefin-12 in the 2016 survey to be 983. This was done to see if performance would be comparable or improved compared to the traditional

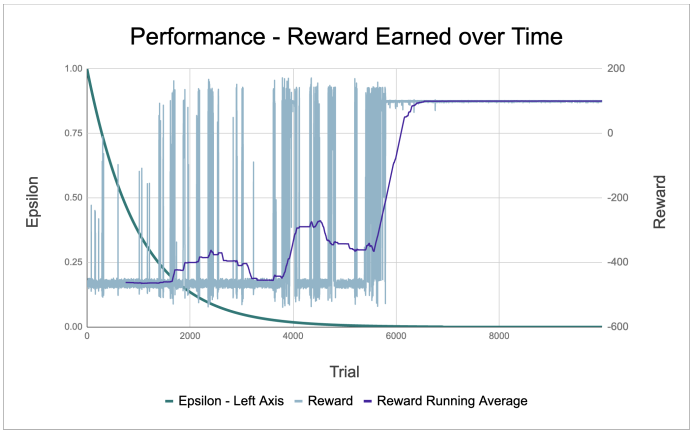

Fig. 9. The left axis depicts the decaying epsilon, greediness of the agent for exploring the environment or exploiting the policy. The epsilon approaches zero around the 6000 trial at which point a dramatic improvement in agent behavior can be seen. The Right Axis depicts the actual reward and the running average reward of the agent, maximum adjusted reward is 100.

localization approach. Upon initiating the survey the agent was able to take the methane and dissolved oxygen data from its state and compute the gradient. The agent descended to the location following the most direct path, resulting in a maximum adjusted reward of 100 and a total number of steps taken to reach the source being 101 including the first move.

B. Discussion

This is an early stage research effort in which we simulated the efficacy of using UMS data to train and deploy an AUV for the discovery of a methane seep in a passive margin pockmark occurring above the hydrate stability zone. The training scenario exhibited a snapshot in time for the distribution of the plume in relation to latitude, longitude, depth and dissolved oxygen. In real-world deployments the environment will be a dynamic stochastic state space which presents more complexity in modeling, training and developing agent behaviors. The presence of methane is known to cause an increase in methane driven decomposition through a multitude of methane oxidation pathways which increase the respiration rate of benthic communities thus driving up the rate of oxygen consumption. This increase in oxygen consumption is estimated to be up to two orders of magnitude higher than normal consumption [21]. These factors of depth and consumption in the presence of methane play a large role in the diffusion and degredation of methane in the environment. Methane diffusion and trace or atmospheric gas interactions are largely time dependent and can be solved as a function of the methane bubbles size and velocity upon ebullition. These metrics are not typically known prior to events and can vary depending on the type of methane emission whether biogenic, thermogenic or pipe burst. For this reason we train the agent on the simplified relationships between the parameters of interest. The potential of integrating additional environmental parameters will allow for learning more comprehensive parameter relationships and target scenarios. Additional environmental parameters could include

salinity, depth, carbon dioxide, nitrogen or other chemical species and may include optical and acoustic measures; which can be pulled from standard sensor packages on-board many commercially available AUV platforms. Other limitations of this work were the richness of data available for simulating the testing environment. Existing methods for analyzing methane distribution via mass spectrometer are either towed systems or deployed on AUV's programmed with predefined trajectories. These pathways are limited in latitude longitude and depth resulting in sparse data. Next steps for this work will integrate computational fluid dynamics for modelling dynamic systems as well as make an effort to partner with data holders to acquire more rich data sets which represent the larger state space for improved simulation of testing environments. The potential of transitioning to a Neural Network (Deep Q-learning) is of interest when evaluating in continuous dynamic environments. This transition would allow the agent to approximate states that have been previously unseen, which is possibly more appropriate for the stochastic nature of this work. It is important to note that the proposed methodology of train, test, verify, update policy based on observations will allow for the development of specialized policies which are catered to an environment, event, and/or region etc. These custom policies have the potential to be shared between agents. Future iterations will test specialized policies within the environment to define an adaptive capability to switch between policies depending on observations. Real-time input from on-board sensing will allow the agent to refine its policy online based on the true in situ observations and possibly swap policies if deemed necessary. It is the goal of this research to progress to real world deployments combining each component of the proposed methodology.

V. Conclusion

In this work we presented a workflow in which an AUV is trained using Q-learning to be an intelligent agent which can locate the source of pollution utilizing chemical signals provided by an on-board UMS. The platform is assumed to be equipped with the developed fluidic sampler for capturing, purifying, labeling and verifying chemical compounds within the test environment via lab-based mass spectrometry analysis. The data observed by both the on-board UMS and the laboratory analysis are used to continue training and improve developed policies for more efficient future deployments. The agent is trained to learn multi-parameter gradient relationships between methane and oxygen for the efficient localization of chemical sources, for this demonstration a biogenic methane seep was identified as the target. The intention of this hybrid training/ground truth approach is to integrate expert knowledge of a specific problem and region into the design of a Q-learning training environment with adequate transferability in order to curate policies that drive more efficient in field mission performance. Given the difficulties in developing high fidelity environmental models which also generalize well for training in different scenarios, a different approach is taken in this effort. An agent is trained using Q-learning in a generic environment with the relevant parameter relationships given the environment and then tested with the learned policy. This real-time chemical sensor and archived sample capture and verification approach yields an adaptive sensing and sampling system. The entire mobile system: AUV mobile platform, reinforcement learning controller, mass spectrometer, and sampler, constitute an adaptive chemical sampling platform.

Acknowledgment

Thank you to Timothy Stewart for your contributions to previous iterations of this work, your efforts are appreciated. We would also like to thank the University of West Florida for the research stipends provided to team members and for the Pensacola Perdido Bay Estuary Program for investing in our research toward adaptive sensing and sampling.

References

[1] P. Stankiewicz, Y. T. Tan, and M. Kobilarov, "Adaptive sampling with an autonomous underwater vehicle in static marine environments," J Field Robotics, vol. 38, no. 4, pp. 572–597, 2021, doi: 10.1002/rob.22005.

[2] Robert W Button, John Kamp, Thomas B Curtin, and James Dryden. A survey of missions for unmanned undersea vehicles. Technical report, RAND NATIONAL DEFENSE RESEARCH INST SANTA MONICA CA, 2009.

[3] D. P. Fries et al., "In-water field analytical technology: Underwater Mass Spectrometer, mobile robots, and remote intelligence for wide and local area chemical profiling," Field Analyt. Chem. Technol., vol. 5, no. 3, pp. 121–130, 2001, doi: 10.1002/fact.1013.

[4] R. T. Short et al., "Underwater mass spectrometers for in situ chemical analysis of the hydrosphere," J Am Soc mass spectrometertrom, vol. 12, no. 6, pp. 676–682, 2001, doi: 10.1016/S1044-0305(01)00246-X.

[5] R. Camilli, B. Bingham, M. Jakuba, H. Singh, and J. Whelan, "Integrating in-situ chemical sampling with auv control systems," in Oceans '04 MTS/IEEE Techno-Ocean '04 (IEEE Cat. No.04CH37600), Kobe, Japan, 2004, vol. 1, pp. 101–109, doi: 10.1109/OCEANS.2004.1402902.

[6] L. Wang, S. Pang, and J. Li, "Olfactory-Based Navigation via Model-Based Reinforcement Learning and Fuzzy Inference Methods," IEEE Trans. Fuzzy Syst., pp. 1–1, 2020, doi: 10.1109/TFUZZ.2020.3011741.

[7] K. Komaki, M. Hatta, K. Okamura, and T. Noguchi, "Development and application of chemical sensors mounting on underwater vehicles to detect hydrothermal plumes," in 2015 IEEE Underwater Technology (UT), Chennai, India, 2015, pp. 1–3, doi: 10.1109/UT.2015.7108248.

[8] K. H. Low, J. Dolan, and P. Khosla, "Information-Theoretic Approach to Efficient Adaptive Path Planning for Mobile Robotic Environmental Sensing," Proceedings of the International Conference on Automated Planning and Scheduling, vol. 19, pp. 233–240, Oct. 2009, Accessed: Aug. 02, 2021. [Online]. Available: https://ojs.aaai.org/index.php/ICAPS/article/view/13344.

[9] M. F. Mysorewala, D. O. Popa, and F. L. Lewis, "Multi-Scale Adaptive Sampling with Mobile Agents for Mapping of Forest Fires," J Intell Robot Syst, vol. 54, no. 4, pp. 535–565, 2009, doi: 10.1007/s10846-008-9246-1.

[10] B. Bayat, N. Crasta, H. Li, and A. Ijspeert, "Optimal search strategies for pollutant source localization," in 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Daejeon, South Korea, 2016, pp. 1801–1807, doi: 10.1109/IROS.2016.7759287.

[11] X. Wu et al., "The autonomous navigation and obstacle avoidance for USVs with ANOA deep reinforcement learning method," Knowledge-Based Systems, vol. 196, p. 105201, 2020, doi: 10.1016/j.knosys.2019.105201.

[12] T. Wiedemann, C. Vlaicu, J. Josifovski, and A. Viseras, "Robotic Information Gathering With Reinforcement Learning Assisted by Domain Knowledge: An Application to Gas Source Localization," IEEE Access, vol. 9, pp. 13159–13172, 2021, doi: 10.1109/ACCESS.2021.3052024.

[13] H. Niu, E. Reeves, and P. Gerstoft, "Source localization in an ocean waveguide using supervised machine learning," J Acoust Soc Am, vol. 142, no. 3, p. 1176, Sep. 2017, doi: 10.1121/1.5000165.

[14] H. Wu, S. Song, K. You, and C. Wu, “Depth Control of Model-Free AUVs via Reinforcement Learning,” IEEE Trans. Syst. Man Cybern, Syst., vol. 49, no. 12, pp. 2499–2510, 2019, doi: 10.1109/TSMC.2017.2785794.

[15] A. C. Sanderson, V. Hombal, D. P. Fries, H. A. Broadbent, J. A. Wilson, P. I. Bhanushali, S. Z. Ivanov, M. Luther, and S. Meyers, “Distributed enviromental sensor network: Design and experiments,” in Multisensor Fusion and Integration for Intelligent Systems, 2006 IEEE International Conference on, pp. 79–84, IEEE, 2006.

[16] D. F. Millie, G. R. Weckman, W. A. Y. II, J. E. Ivey, D. P. Fries, E. Ardjmand, and G. L. Fahnenstiel, “Coastal ’big data’ and nature-inspired computation: Prediction potentials, uncertainties, and knowledge derivation of neural networks for an algal metric,” Estuarine, Coastal and Shelf Science, vol. 125, no. 0, pp. 57 – 67, 2013.

[17] D. Fries, G. Barton, G. Hendrick, B. Gregson, L. Hotaling, J. Paul, A. Sanderson, and R. Blidberg, “Solar robotic material sampler system for chemical, biological and physical ocean observations,” in OCEANS 2011, pp. 1–5, IEEE, 2011.

[18] Mau, S., Römer, M., Torres, M. et al. Widespread methane seepage along the continental margin off Svalbard - from Bjørnøya to Kongsfjorden. Sci Rep 7, 42997 (2017). https://doi.org/10.1038/srep42997

[19] T. Gentz, E. Damm, J. Schneider von Deimling, S. Mau, D. F. McGinnis, and M. Schlüter, “A water column study of methane around gas flares located at the West Spitsbergen continental margin,” Continental Shelf Research, vol. 72, pp. 107–118, 2014, doi: 10.1016/j.csr.2013.07.013.

[20] K. J. Cunningham and R. Westcott, “Methane, nitrogen, oxygen, and carbon dioxide data for the Miami Pockmark and Key Biscayne Pockmark.” U.S. Geological Survey, 2020, doi: 10.5066/F7H13196.

[21] A. Boetius and F. Wenzhöfer, “Seafloor oxygen consumption fuelled by methane from cold seeps,” Nature Geosci, vol. 6, no. 9, pp. 725–734, Sep. 2013, doi: 10.1038/ngeo1926.

[22] H. A. Broadbent, S. Z. Ivanov and D. P. Fries, ”PCB-MEMS Environmental Sensors in the Field,” 2007 IEEE International Symposium on Industrial Electronics, 2007, pp. 3282-3286, doi: 10.1109/ISIE.2007.4375141.