# Information-Guided Robotic Maximum Seek-and-Sample in Partially Observable Continuous Environments

Genevieve Flaspohler*[1,2] Victoria Preston*[1,2] Anna P.M. Michel[2] Yogesh Girdhar[2] Nicholas Roy[1]

*Abstract*—We present PLUMES, a planner for localizing and collecting samples at the global maximum of an *a priori* unknown and partially observable continuous environment. This "maximum seek-and-sample" (MSS) problem is pervasive in the environmental and earth sciences. Experts want to collect scientifically valuable samples at an environmental maximum (e.g., an oil-spill source), but do not have prior knowledge about the phenomenon's distribution. We formulate the MSS problem as a partially-observable Markov decision process (POMDP) with continuous state and observation spaces, and a sparse reward signal. To solve the MSS POMDP, PLUMES uses an information-theoretic reward heuristic with continuous-observation Monte Carlo Tree Search to efficiently localize and sample from the global maximum. In simulation and field experiments, PLUMES collects more scientifically valuable samples than state-of-the-art planners in a diverse set of environments, with various platforms, sensors, and challenging real-world conditions.

*Index Terms*—Motion and Path Planning, Learning and Adaptive Systems, Reactive and Sensor-Based Planning

## I. INTRODUCTION

IN many environmental and earth science applications, experts want to collect scientifically valuable samples of a maximum (e.g., an oil spill source), but the distribution of the phenomenon is initially unknown. This *maximum seek-and-sample* (MSS) problem is pervasive. Canonically, samples are collected at predetermined locations by a technician or by a mobile platform following a uniform coverage trajectory. These non-adaptive strategies result in sample sparsity at the maximum and may be infeasible when the geometric structure of the environment is unknown (e.g., boulder fields) or changing (e.g., tidal zones). Increasing the number of valuable samples at the maximum requires adaptive online planning and execution. We present PLUMES — **P**lume **L**ocalization under **U**ncertainty using **M**aximum-Valu**E** information and **S**earch — an adaptive algorithm that enables a mobile robot to efficiently localize and densely sample an environmental maximum, subject to practical challenges including dynamic
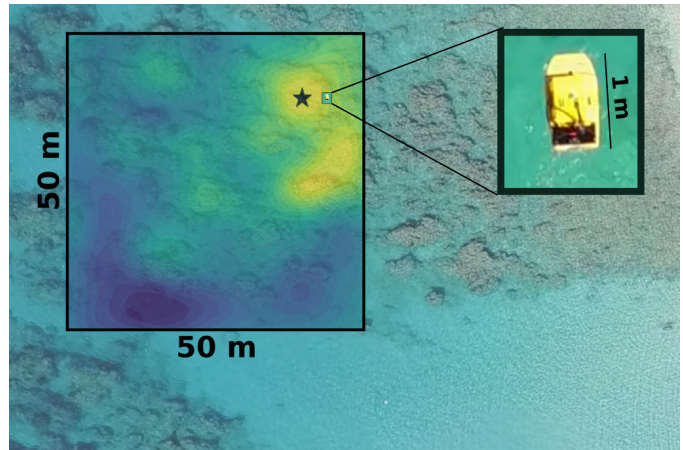
Fig. 1. **Coral head localization with an autonomous surface vehicle (ASV):** The objective of the ASV is to find and sample at the most exposed (shallowest) coral head in a region of Bellairs Fringing Reef, Barbados. Overlaid on the aerial photo is the *a priori* unknown bathymetry of the region (yellow is shallow, blue is deep). Equipped with an acoustic point altimeter, the ASV must explore to infer the location of the maximum (marked with a star) and then sample at that coral colony.

constraints, unknown geometric map and obstacles, and noisy sensors with limited field-of-view. Fig. 1 shows a motivating application of coral head localization.

**Informative Path Planning:** The MSS problem is closely related to informative path planning (IPP) problems. Offline IPP techniques for pure information-gathering that optimize submodular coverage objectives can achieve near-optimal performance [1], [2]. However, in the MSS problem, the value of a sample depends on the unknown maximum location, requiring a robot to adaptively select actions that explore to localize the maximum and then seamlessly transition to selecting actions that exploitatively collect valuable samples there. Even for adaptive IPP methods, the MSS problem presents considerable challenges. The target environmental phenomenon is partially observable and most directly modeled as a continuous scalar function. Additionally, efficient maximum sampling with a mobile robot requires consideration of vehicle dynamics, travel cost, and a potentially unknown obstacle map. Handling these challenges in combination excludes adaptive IPP algorithms that use discrete state spaces [3], [4], known metric maps [5], [6], or unconstrained sensor placement [7].

**The MSS POMDP:** Partially-observable Markov decision processes (POMDPs) are general models for decision-making under uncertainty that allow the challenging aspects of the

MSS problem to be encoded. We define the MSS POMDP in which the partially observable state represents the continuous environmental phenomenon and a sparse reward function encodes the MSS scientific objective by giving reward only to samples sufficiently close to the global maximum. Solving a POMDP exactly is generally intractable, and the MSS POMDP is additionally complicated by both continuous state and observation spaces, and the sparse MSS reward function. This presents the two core challenges that PLUMES addresses: performing online search in a belief-space over continuous functions, and overcoming reward function sparsity.

**Planning over Continuous Domains:** In the MSS problem, the state of the environment can be modeled as a continuous function. PLUMES uses a Gaussian Process (GP) model to represent the belief over this continuous function, and must plan over the uncountable set of possible GP beliefs that arise from future continuous observations. To address planning in continuous spaces, state-of-the-art online POMDP solvers use deterministic discretization [8] or a combination of sampling techniques and particle filter belief representations [9]–[12]. Efficiently discretizing or maintaining a sufficiently rich particle set to represent the underlying continuous function in MSS applications is itself a challenging problem, and can lead to inaccurate inference of the maximum [13]. Other approaches have considered using the maximum-likelihood observation to make search tractable [14]. However, this assumption can compromise search and has optimality guarantees only in linear-Gaussian systems [15]. Instead, PLUMES uses Monte Carlo Tree Search (MCTS) with progressive widening to limit planning tree growth [16] and retains optimality guarantees [17] in continuous environments.

**Rewards and Heuristics:** In the MSS POMDP, the reward function is sparse and does not explicitly encode the value of exploration. Planning with sparse rewards requires long-horizon information gathering and is an open problem in robotics [18]. To alleviate this difficulty, less sparse heuristic reward functions can be optimized in place of the true reward, but these heuristics need to be selected carefully to ensure the planner performs well with respect to the true objective. In IPP, heuristics based on the value of information have been applied successfully [7], [14], [19], [20], primarily using the GP-UCB criteria [1], [21]. We demonstrate that within practical mission constraints, using UCB as the heuristic reward function for the MSS POMDP can lead to suboptimal convergence to local maxima due to a mismatch between the UCB heuristic and the true MSS reward. Instead, PLUMES takes advantage of a heuristic function from the Bayesian optimization (BO) community for state-of-the-art black-box optimization [22] which we call *maximum-value information* (MVI). MVI overcomes sparsity and encourages long-term information gathering, while still converging to the true reward of the MSS POMDP.

The contribution of this paper is the MSS POMDP formalism and the corresponding PLUMES planner, which by virtue of its belief model, information-theoretic reward heuristic, and search framework, enables efficient maximum sampling with asymptotic optimality guarantees for continuous environments. PLUMES extends the state-of-the-art in MSS planners by applying a BO heuristic reward function to MSS that alleviates the challenges of the true sparse MSS reward function, and integrating GP belief representations within continuous-observation MCTS. The utility of PLUMES for MSS applications is demonstrated in extensive simulation and field trials, showing a statistically significant performance improvement over state-of-the-art baselines.

## II. MAXIMUM SEEK-AND-SAMPLE POMDP

We formalize the MSS problem by considering a target environmental domain as a $d$-dimensional compact set $\mathbb{X}_w \subset \mathbb{R}^d$. We allow $\mathbb{X}_w$ to contain obstacles with arbitrary geometry and let $\mathbb{X} \subset \mathbb{X}_w$ be the set of reachable points with respect to the robot's initial pose. We assume there is an unknown underlying continuous function $f : \mathbb{X}_w \to \mathbb{R}$ representing the value of a continuous phenomenon of interest. The objective is to find the unique global maximizer $\mathbf{x}^* = \arg\max_{\mathbf{x} \in \mathbb{X}} f(\mathbf{x})$ by safely navigating while receiving noisy observations of this function $f$. Because $f$ is unknown, we cannot access derivative information or any analytic form.

We model the process of navigating and generating observations as the MSS POMDP: an 8-tuple $(\mathcal{S}, \mathcal{A}, \mathcal{Z}, T, O, R, \gamma, b_0)$:

- $\mathcal{S}$: continuous state space of the robot and environment
- $\mathcal{A}$: discrete set of action primitives
- $\mathcal{Z}$: continuous space of possible observations
- $T: \mathcal{S} \times \mathcal{A} \to \mathcal{P}(\mathcal{S})$, the transition function, i.e., $\mathbf{Pr}(S_{t+1} = s' \mid S_t = s, A_t = a)$
- $O: \mathcal{S} \times \mathcal{A} \to \mathcal{P}(\mathcal{Z})$, the observation model, i.e., $\mathbf{Pr}(Z_{t+1} = z \mid S_{t+1} = s, A_t = a)$
- $R: \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, the reward of taking action $a$ when robot's state is $s$, i.e., $R(s, a)$
- $\gamma$: discount factor, $0 \leq \gamma \leq 1$
- $b_0$: initial belief state of the robot, $b_0 \in \mathcal{P}(S_0)$

where $\mathcal{P}(\cdot)$ denotes the space of probability distributions over the argument.

The Bellman equation is used to recursively quantify the value of belief $b_t = \mathcal{P}(S_t)$ over a finite horizon $h$ under policy $\pi : b_t \to a_t$ as:

$$V_h^\pi(b_t) = \mathbb{E}[R(s_t, \pi(b_t))] + \\ \gamma \int_{z \in \mathcal{Z}} V_{h-1}^\pi(b_{t+1}^{\pi(b_t),z}) \, \mathbf{Pr}(z \mid b_t, \pi(b_t)) \, \mathrm{d}z, \quad (1)$$

where the expectation is taken over the current belief and $b_{t+1}^{\pi(b_t),z}$ is the updated belief after taking action $\pi(b_t)$ and observing $z \in \mathcal{Z}$. The optimal policy $\pi_h^*$ over horizon-$h$ is the maximizer of the value function over the space of possible policies $\Pi$: $\pi_h^* = \arg\max_{\pi \in \Pi} V_h^\pi(b_t)$. However, Eq. 1 is intractable to compute in continuous state and observation spaces; the optimal policy must be approximated. PLUMES uses a receding-horizon, online POMDP planner and heuristic reward function to approximately solve the MSS POMDP in real-time on robotic systems.

## III. THE PLUMES ALGORITHM

PLUMES is an online planning algorithm with a sequential decision-making structure:

1) Conditioned on $b_t$, approximate the optimal policy $\pi_h^*$ for finite horizon $h$ and execute the action $a = \hat{\pi}_h^*(b_t)$.

2) Collect observations $z \in \mathcal{Z}$, according to $O$.

3) Update $b_t$ to incorporate this new observation; repeat.

In the following sections, we define the specific choice of belief model, search algorithm, and heuristic reward function PLUMES uses to solve the MSS POMDP.

### A. Gaussian Process Belief Model

We assume the robot's pose $\mathbf{x}_t$ at planning iteration $t$ is fully observable, and the unknown environmental phenomenon $f$ is partially observable. The full belief-state is represented as a tuple $b_t$ of robot state $\mathbf{x}_t$ and environment belief $g_t = \mathcal{P}(f)$ at time $t$. Because $f$ is a continuous function, we cannot represent the belief $g_t$ as a distribution over discrete states, as is standard in POMDP literature [23], and must choose an alternate representation. PLUMES uses a Gaussian process (GP) [24] to represent $g_t$ conditioned on a history of past observations. This GP is parameterized by mean $\mu(\mathbf{x})$ and covariance function $\kappa(\mathbf{x}, \mathbf{x}')$.

As the robot traverses a location $\mathbf{x}$, it gathers observations $z \in \mathcal{Z}$ of $f$ subject to sensor noise $\sigma_n^2$, such that $z = f(\mathbf{x}) + \epsilon$ with $\epsilon \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_n^2)$. Given a history $\mathcal{D}_t = \{\mathbf{x}_i, z_i\}_{i=0}^D$ of $D$ observations and observation locations at planning iteration $t$, the posterior belief at a new location $\mathbf{x}' \in \mathbb{X}$ is computed:

$$g_t(\mathbf{x}') \mid \mathcal{D}_t \sim \mathcal{N}(\mu_t(\mathbf{x}'), \sigma_t^2(\mathbf{x}')), \text{where} \tag{2}$$

$$\mu_t(\mathbf{x}') = \kappa_t(\mathbf{x}')^\top (\mathbf{K}_t + \sigma_n^2 \mathbf{I})^{-1} \mathbf{z}_t, \tag{3}$$

$$\sigma_t^2(\mathbf{x}') = \kappa(\mathbf{x}', \mathbf{x}') - \kappa_t(\mathbf{x}')^\top (\mathbf{K}_t + \sigma_n^2 \mathbf{I})^{-1} \kappa_t(\mathbf{x}'), \tag{4}$$

where $\mathbf{z}_t = [z_0, \ldots, z_{D-1}]^\top$, $\mathbf{K}_t$ is the positive definite kernel matrix with $\mathbf{K}_t[i, j] = \kappa(\mathbf{x}_i, \mathbf{x}_j)$ for all $\mathbf{x}_i, \mathbf{x}_j \in \mathcal{D}_t$, and $\kappa_t(\mathbf{x}') = [\kappa(\mathbf{x}_0, \mathbf{x}'), \ldots, \kappa(\mathbf{x}_{D-1}, \mathbf{x}')]^\top$.

### B. Planning with Continuous-Observation MCTS

PLUMES selects high-reward actions with receding-horizon search over possible belief states. This search requires a simulator that can sample observations and generate beliefs given a proposed action sequence. For PLUMES, this simulator is the GP model, which represents the belief over the continuous function $f$, and in turn generates continuous observations from action queries using Eq. 3 & 4.

PLUMES uses continuous-observation MCTS to overcome the challenges of planning in continuous state and observation spaces. There are three stages: *selection*, *forward simulation*, and *back-propagation*. Each node in the tree can be represented as the tuple of robot pose and GP belief, $b_t = \{\mathbf{x}_t, g_t\}$. Additionally, we will refer to two types of nodes: belief nodes and belief-action nodes. The root of the tree is always a belief node, which represents the entire history of actions and observations up through the current planning iteration. Through selection and simulation, belief and belief-action nodes are alternately added to the tree (Fig. 2).

From the root, a rollout begins with the *selection* stage, in which a belief-action child is selected according to the Polynomial Upper Confidence Tree (PUCT) policy [17]. The PUCT value $\hat{Q}_{aug}^*(b_t, a)$ is the sum of the average heuristic
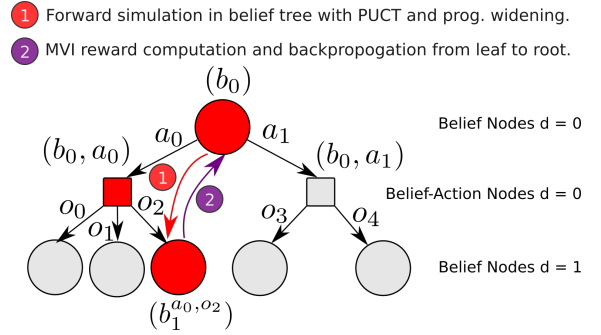


Fig. 2. **Continuous-observation MCTS:** Illustrated to horizon $h = 1$, the tree consists of alternating belief and belief-action nodes. Action decisions are made at belief nodes and random belief transitions according to the observation function occur at belief-action nodes. Note that belief-action nodes have a varying number of children due to progressive widening and unequal simulation (not visualized) due to PUCT policy.

rewards (i.e., MVI) from all previous simulations and a term that favors less-simulated action sequences:

$$\hat{Q}_{aug}^*(b_t, a) = \hat{Q}^*(b_t, a) + \sqrt{\frac{N(b_t)^{e_d}}{N(b_t, a)}}, \tag{5}$$

where $\hat{Q}^*(b_t, a)$ is the average heuristic reward of choosing action $a$ with belief $b_t$ in all previous rollouts, $N(b_t)$ is the number of times the node $b_t$ has been simulated, $N(b_t, a)$ is the number of times that particular action from node $b_t$ has been selected, and $e_d$ is a depth-dependent parameter*.

Once a child belief-action node is selected, the action associated with the child is *forward simulated* using the generative observation model $O$, and a new belief node is generated $b_{t+1} = \{\mathbf{x}_{t+1}, g_{t+1}\}$ as though the action were taken and samples observed. The simulated observations are drawn from the belief-action node's GP model $g_t$, and the robot's pose is updated deterministically based on the selected action. Since the observations in a GP are continuous, every sampled observation is unique with probability one. *Progressive widening*, with depth-dependent parameter* $\alpha_d$ incrementally grows the tree by limiting the number of belief children of each belief-action node. When growing the tree, $b_{t+1}$ is either chosen to be the least visited node if $\lfloor N(b_t, a)^{\alpha_d} \rfloor = \lfloor (N(b_t, a) - 1)^{\alpha_d} \rfloor$, or otherwise is a new child with observations simulated from $b_t$. By limiting the width of the search tree and incrementally growing the number of explored children, progressive widening avoids search degeneracy in continuous environments.

Once a sequence of actions has been rolled out to a horizon $h$, the accumulated heuristic reward is *propagated upward* from the leaves to the tree root. The average accumulated heuristic reward and number of queries are updated for each node visited in the rollout. Rollouts continue until the computation budget is exhausted. The most visited belief-action child of the root node is executed.

Continuous-observation MCTS within PLUMES provides both practical and theoretical benefits. Practically, progressive-widening directly addresses search degeneracy by visiting belief nodes multiple times even in continuous observation spaces, allowing for a more representative estimate of their

*Refer to Table 1 of Auger et al. [17] for parameter settings.

value. Theoretically, PLUMES can be shown to select asymptotically optimal actions. We briefly describe how analysis in Auger et al. [17] for PUCT-MCTS with progressive widening in MDPs can be extended to PLUMES.

Using standard methods [23], we can reduce the MSS POMDP to an equivalent belief-state MDP. This belief-state MDP has a state space equal to the set of all possible beliefs, and a transition distribution that captures the effect of both the dynamics and the observation model after each action. Planning in this representation is often intractable as the state space is continuous and infinite-dimensional. However, PLUMES plans directly in the belief-state MDP by using its GP belief state to compute the transition function efficiently.

Subsequently, Theorem 1 in Auger et al. [17] shows that for an MDP with a continuous state space, like the belief-state MDP representation suggested, the value function estimated by continuous-observation MCTS asymptotically converges to that of the optimal policy:

$$\left| \hat{Q}_h^*(b_t, a) - Q_h^*(b_t, a) \right| \leq \frac{C}{N(b_t, a)^{\gamma_d}}, \tag{6}$$

with high probability [17], for constants $C > 0$ and $\gamma_{d*}$.

### C. Maximum-Value Information Reward

The true state-dependent reward function for the MSS POMDP would place value on collecting sample points $\mathbf{x}$ within an $\epsilon$-ball of the true global maximum $\mathbf{x}^*$:

$$R(f, \mathbf{x}) = \mathbb{1}_{\|\mathbf{x} - \mathbf{x}^*\| < \epsilon}, \tag{7}$$

where $\epsilon$ is determined by the scientific application. Optimizing this sparse reward function directly is challenging, so PLUMES approximates the true MSS reward by using the maximum-value information (MVI) heuristic reward [22].

MVI initially encourages exploration behavior, but ultimately rewards exploitative sampling near the inferred maximum.

The belief-dependent MVI heuristic reward $\tilde{R}(b_t, \mathbf{x})$ quantifies the expected value of having belief $b_t$ and collecting a sample at location $\mathbf{x} \in \mathbb{X}$. MVI reward quantifies the mutual information between the random variable $Z$, representing the observation at location $\mathbf{x}$, and $Z^*$, the random variable representing the value of the function $f$ at the global maximum:

$$\tilde{R}(b_t, \mathbf{x}) = I(\{\mathbf{x}, Z\}; Z^* \mid b_t), \tag{8}$$

where $Z^* = \max_{\mathbf{x}' \in \mathbb{X}} f(\mathbf{x}')$. To compute the reward of collecting a random observation $Z$ at location $\mathbf{x}$ under belief $b_t$, we approximate the expectation over the unknown $Z^*$ by sampling from the posterior distribution $z_i^* \sim p(Z^* \mid b_t)$ and use Monte Carlo integration with $M$ samples [22]:

$$\tilde{R}(b_t, \mathbf{x}) = H[\mathbf{Pr}(Z \mid \mathbf{x}, b_t)] -$$
$$\mathbb{E}_{z' \sim \mathbf{Pr}(Z^* \mid b_t)}[H[\mathbf{Pr}(Z \mid \mathbf{x}, b_t, Z^* = z')], \tag{9}$$

$$\approx H[\mathbf{Pr}(Z \mid \mathbf{x}, b_t)] - \frac{1}{M}\sum_{i=0}^{M} H[\mathbf{Pr}(Z \mid \mathbf{x}, b_t, Z^* = z_i^*)]. \tag{10}$$

Each entropy expression $H[\cdot]$ can be respectively approximated as the entropy of a Gaussian random variable with mean and variance given by the GP equations (Eq. 3 & 4), and the entropy of a truncated Gaussian, with upper limit $z_i^*$ and the same mean and variance.

To draw samples $z_i^*$ from the posterior $p(Z^* \mid b_t)$, we employ spectral sampling [25]. Spectral sampling draws a function $\hat{f}$, which has analytic form and is differentiable, from the posterior belief of a GP with stationary covariance function [22], [26]. To complete the evaluation of Eq. 10, $z_i^* \sim p(Z^* \mid b_t)$ can be computed by applying standard efficient global optimization techniques (e.g., sequential least
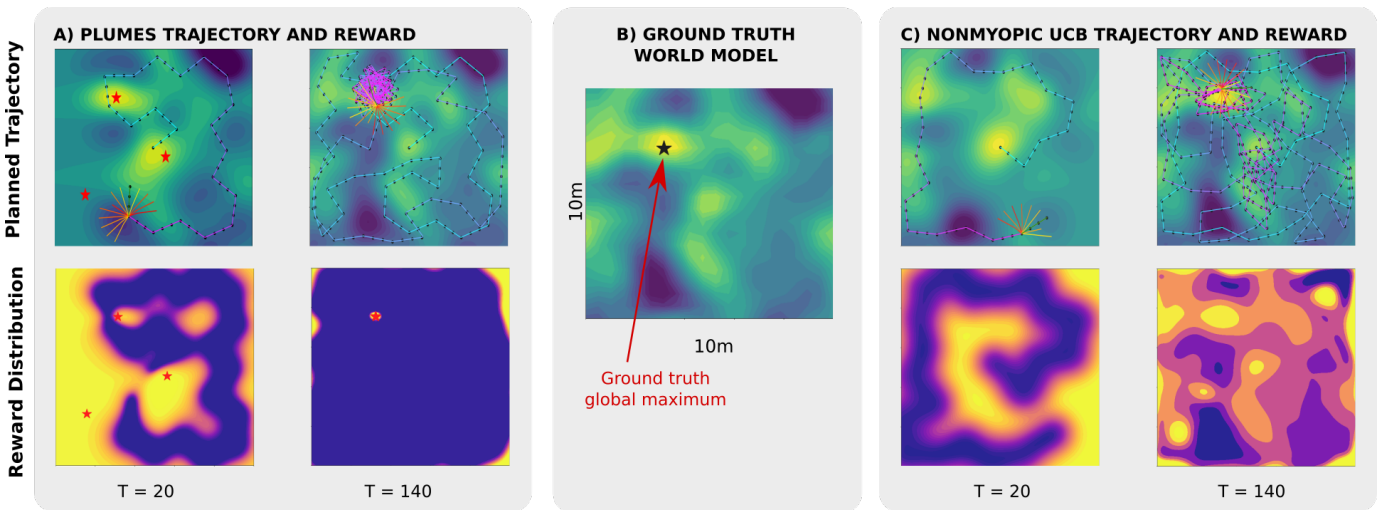


Fig. 3. **Convergence of MVI vs UCB heuristic:** The true environmental phenomenon with the global maximum marked by a star is shown in the center; high regions are colored yellow and low regions blue. In (A,C), the robot trajectory and corresponding reward functions are shown early (20 actions) and later (140 actions) in a mission. On the top row, snapshots of the robot belief state with planned trajectories are shown, with recent actions colored pink and earlier actions colored blue. Red stars mark maxima sampled by MVI. In the bottom row, the corresponding reward function is shown, with high-reward regions colored yellow and low reward regions colored purple. By the end of the mission, MVI clearly converges to placing reward only at the global maximum, which in turn leads to efficient convergence of the robot. By contrast, the reward landscape resulting from canonically used UCB converges to the underlying function, causing the UCB planner to uniformly tour high-valued regions of the environment.

squares programming, quasi-Newton methods) to find the global maximum of the sampled $\hat{f}$. This results in the following expression for MVI reward [22]:

$$\tilde{R}(b_t, \mathbf{x}) \approx \frac{1}{M} \sum_{i=0}^{M} \frac{\gamma_{z_i^*}(\mathbf{x})\phi(\gamma_{z_i^*}(\mathbf{x}))}{2\Phi(\gamma_{z_i^*}(\mathbf{x}))} - \log(\Phi(\gamma_{z_i^*}(\mathbf{x}))) \quad (11)$$

where $\gamma_{z_i^*}(\mathbf{x}) = \frac{z_i^* - \mu_t(\mathbf{x})}{\sigma_t(\mathbf{x})}$, $\mu_t(x)$ and $\sigma_t(x)$ are given by Eq. 3 & 4, and $\phi$ and $\Phi$ are the standard normal PDF and CDF. For actions that collect samples at more then one location, the reward of an action $\tilde{R}(b_t, a)$ is the sum of rewards of the locations sampled by that action.

MVI initially favors collecting observations in areas that have high uncertainty due to sampling maxima from the initial uniform GP belief. As observations are collected and uncertainty diminishes in the GP, the sampled maxima converge to the true maximum and reward concentrates locally at this point, encouraging exploitative behavior. This contrasts with the Upper Confidence Bound (UCB) heuristic, which distributes reward proportional to predictive mean $\mu_t(\mathbf{x})$ and weighted variance $\sigma_t(\mathbf{x})$ of the current GP belief model (Eq. 3 & 4): $\tilde{R}(b_t, \mathbf{x}) = \mu_t(\mathbf{x}) + \sqrt{\beta_t}\sigma(\mathbf{x})$. As the robot explores, UCB converges to the underlying phenomenon, $f$. The difference in convergence characteristics between MVI and UCB can be observed in Fig. 3.

## IV. EXPERIMENTS AND RESULTS

We analyze the empirical performance of PLUMES in a breadth of MSS scenarios that feature convex and non-convex environments. We compare against three baselines used in environmental surveying: non-adaptive lawnmower-coverage (Boustro., an abbreviation of boustrophedonic [27]), greedy myopic planning with UCB reward (UCB-Myopic) [19], and nonmyopic planning with traditional MCTS [28] that uses the maximum-likelihood observation and UCB reward (UCB-MCTS) [14]. The performance of UCB planners has been shown to be sensitive with respect to $\beta$ value [14]. In order to avoid subjective tuning, we select a time-varying $\beta_t$ that is known to enable no-regret UCB planning [1], [19]. PLUMES uses continuous-observation MCTS with hyperparameters presented in Auger et al. [17].

To evaluate the mission performance of all planners, we report accumulated MSS reward (Eq. 7), which directly corresponds to the number of scientifically valuable samples collected within an $\epsilon$-ball of the true maximum. This metric is reported for all trial scenarios in Table I. We additionally report several metrics commonly used in IPP to evaluate posterior model quality: overall environmental posterior root mean-squared error (RMSE) and error in posterior prediction of $\mathbf{x}^*$ at the end of a mission ($\mathbf{x}^*$ error). We use a Mann-Whitney U non-parametric significance test [29] to report statistical significance (p = 0.05 level) in performance between PLUMES and baseline algorithms.

### A. Bounded Convex Environments

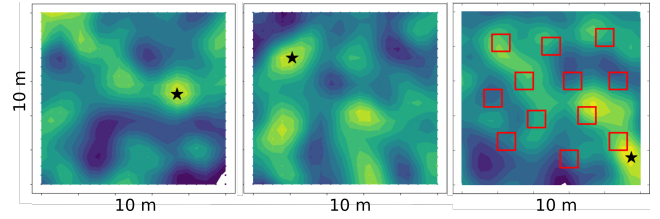In marine and atmospheric applications, MSS often occurs in a geographically bounded, obstacle-free environment. In



Fig. 4. **Simulation Environments:** The multimodal simulated 10m × 10m environments. Yellow regions are high-valued; blue regions are low-valued. The global maximum is marked with a star. The left and center environments represent convex-worlds (Section IV-A), while the right environment is representative of a non-convex world (Section IV-B).

50 simulated trials, we applied PLUMES and our baseline planners to a point robot in a $10\,\text{m} \times 10\,\text{m}$ multimodal environment drawn randomly from a GP prior with a squared-exponential covariance function and zero mean ($l = 1.0$, $\sigma^2 = 100.0$, $\sigma_n^2 = 1.0$ [1%]) (see Fig.4). The action set consisted of ten viable trajectories centered at the robot's pose with path length $1.5\,\text{m}$, and samples are collected every $0.5\,\text{m}$ of travel. Mission lengths were budgeted to be $200\,\text{m}$. Nonmyopic planners rolled out to a 5-action horizon and were allowed 250 rollouts per planning iteration. Summary simulation results are presented in Table I.

In these trials, PLUMES accumulated significantly (0.05-level) more reward than baselines. The distribution of accumulated reward (Fig. 5) shows that PLUMES has a single dominating mode near reward 200 and few low-performing missions (reward <50). In contrast, both UCB-based methods have distributions which are multimodal, with non-trivial modes in the low-performance region. Boustro. collected consistently few scientifically valuable samples. In addition to collecting many more samples at the maximum, PLUMES achieved statistically indistinguishable levels of posterior RMSE and $\mathbf{x}^*$ error compared to baselines (Table I).

The corresponding field trial for convex-world maximum-search was performed in the Bellairs Fringing Reef, Barbados
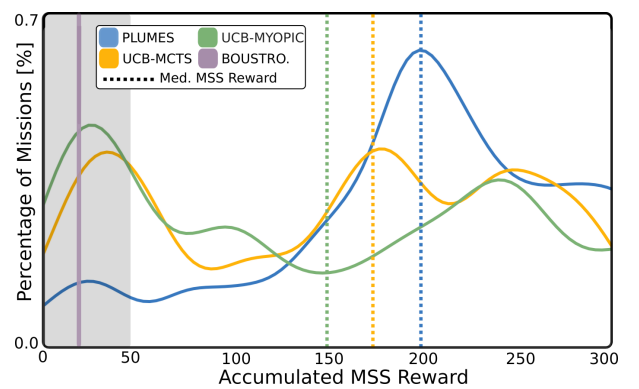


Fig. 5. **Distribution of accumulated MSS reward in 50 convex-world simulations:** Accumulated MSS reward is calculated for each trial and the distribution for each planner is plotted as a kernel density estimate (solid line). The dashed lines represent the median accumulated reward for each planner (reported in Table I). The gray area of the plot indicates a low performance region where the planner collected <50 samples near the maximum. PLUMES has a single mode near 200, whereas both UCB-based methods are multi-modal, with modes in the low performance region.

TABLE I
ACCUMULATED TRUE MSS REWARD (EQ. 7), RMSE, AND $\mathbf{x}^*$ ERROR, REPORTED AS MEDIAN (INTERQUARTILE RANGE).
ASTERISKS DENOTE BASELINES WHOSE DIFFERENCE IN PERFORMANCE IS STATISTICALLY SIGNIFICANT COMPARED TO PLUMES.

| | Convex Simulation Trials $\epsilon = 1.5$ m, 50 trials | | | ASV Trial $\epsilon = 10$ m, 1 trial | Non-convex Simulation Trials $\epsilon = 1.5$ m, 50 trials | | | Dubins Car Trials $\epsilon = 1.5$ m, 5 trials |
| | MSS Reward | RMSE | $\mathbf{x}^*$ Error | MSS Reward | MSS Reward | RMSE | $\mathbf{x}^*$ Error | MSS Reward |
|---|---|---|---|---|---|---|---|---|
| PLUMES | **199** (89) | 3.8 (9.2) | 0.21 (0.23) | **524** | **206** (100) | 3.6 (2.1) | 0.25 (0.56) | **159** (74) |
| UCB-MCTS | 171 (179)* | 3.7 (9.6) | 0.24 (0.29) | - | 115 (184)* | 3.6 (1.5) | 0.27 (1.18) | 52 (17) |
| UCB-Myopic | 148 (199)* | 3.6 (9.2) | 0.33 (3.25) | - | 86 (102)* | 3.4 (1.0) | 0.23 (0.34) | 42 (66) |
| Boustro. | 27 (3)* | 2.7 (10.4) | 0.26 (0.46) | 63 | - | - | - | - |

by a custom-built autonomous surface vehicle (ASV) with the objective of localizing the most exposed coral head. Coral head exposure is used to select vantage points for coral imaging [30] and in ultraviolet radiation studies on coral organisms [31]. Due to time and resource constraints, only one trial of two planners was feasible on the physical reef; we elected to demonstrate PLUMES and Boustro., one of the most canonical surveying strategies in marine sciences.

The ASV ($1$ m $\times$ $0.5$ m) had holonomic dynamics and a downward-facing acoustic point altimeter (Tritech Micron Echosounder) with returns at $1$ Hz. Ten dynamically-feasible $10$ m straight paths radiating from the location of the ASV were used in the action set. The environment was bounded by a $50$ m by $50$ m geofence. Localization and control was provided by a PixHawk Autopilot with GPS and internal IMU; the fused state estimate was empirically suitable for the desired maximum localization accuracy ($\epsilon = 10$ m). The budget for each mission was $1000$ m, which took approx. 45 minutes to travel. The GP kernel was trained on altimeter data from a dense data collection deployment the day before (parameters $l = 2.01$, $\sigma^2 = 0.53$, $\sigma_n^2 = 0.02$ [26%]). Note the high noise in the inferred GP model, as well as the relatively small length-scale in the $2500$ m$^2$ field site. The reconstructed bathymetry and vehicle are shown in Fig. 6.

PLUMES successfully identified the same coral head to be maximal as that inferred from the GP trained on prior dense data collection, as indicated by accumulated reward in Table I, overcoming the challenges of moving in ocean waves, noisy altimeter measurements, and highly multimodal environment. Additionally, the posterior prediction of $\mathbf{x}^*$ had an error of only $1.78$ m while Boustro. reported $8.75$ m error due to its



Fig. 7. **Extending PLUMES for Spatiotemporal Monitoring:** (A) The ground truth map at two planning iterations for a dynamic environment. The maximum is marked with a black star, and migrates from the top left to the top right of the world. (B) MVI reward is redistributed by using a spacetime kernel within PLUMES that captures the environment's dynamics.

non-adaptive sampling strategy.

In the Bellairs Fringing Reef trials, the environment was assumed to be static. However, in many marine domains the impact of sediment transport, waves, and tides could physically change the location of a maximum over the course of a mission. PLUMES can be extended to dynamic environments by employing a spatiotemporal kernel in the GP model, which allows for the predictive mean and variance to change temporally [32]. If the dynamics of an environment can be encoded in the kernel function, no other changes to PLUMES are necessary; MVI will be distributed according to the time dynamic. Fig. 7 demonstrates the properties of PLUMES with a squared-exponential kernel over space ($l = 1.5$, $\sigma^2 = 100$, $\sigma_n^2 = 0.5$) and time ($l = 100$, $\sigma^2 = 100$, $\sigma_n^2 = 0.5$). In this illustrative scenario, the global maximum moved between planning iteration $T = 230$ and $T = 250$. PLUMES with a spatiotemporal kernel maintained multiple hypotheses about the maximum's location given the random-walk dynamic of the environment, resulting in MVI reward being re-distributed between the two maxima over time.

### B. Non-Convex Environments

We next consider non-convex environments with potentially unknown obstacles, a situation that occurs frequently in practical MSS applications with geographical no-go zones for rover or ASV missions, and in indoor or urban settings. We evaluated PLUMES, UCB-Myopic, and UCB-MCTS planners
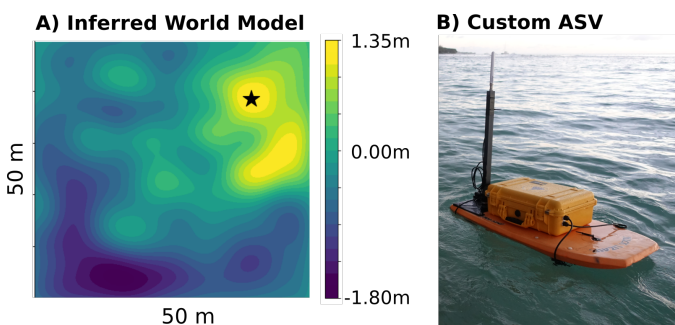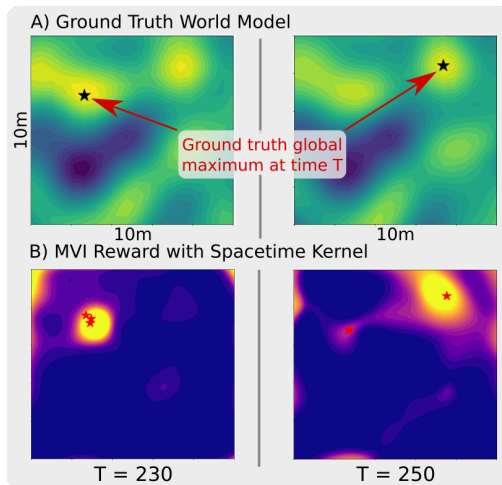


Fig. 6. **Coral head map and ASV:** (A) The ground truth bathymetric map inferred from all collected data, mean corrected in depth. Yellow represents shallower depths, and blue is deeper. The global maximum is marked with a black star. (B) The custom ASV used to traverse the $2500$ m$^2$ region.
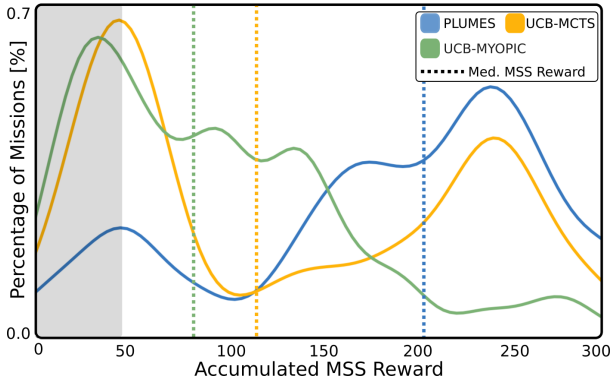
Fig. 8. **Distribution of accumulated MSS reward in 50 non-convex mission simulations:** Accumulated MSS reward distribution (solid line) and median (dashed line, reported in Table I) for each planner. The gray area of the plot indicates a low performance region (reward <50). PLUMES has few low-performing missions and a primary mode near reward 250. The primary mode of both UCB-based methods is in the low performance region due to convergence to suboptimal local maxima.
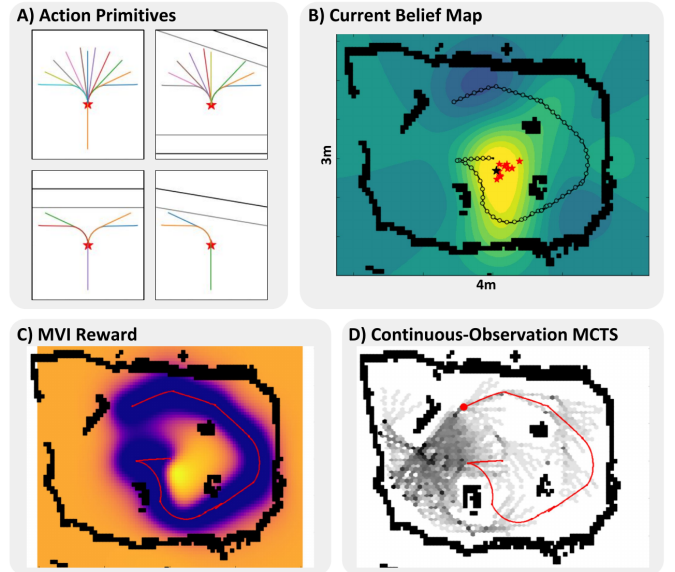


Fig. 9. **Snapshot of unknown non-convex map scenario:** (A) shows examples of how the action-primitives change based upon obstacle detection (black lines) and safety padding (grey lines). (B-D) show a planning iteration of PLUMES, starting with the current belief map and obstacle detections (B). The MVI heuristic is illustrated in (C) where lighter regions are higher value. (D) shows the rollout visibility of continuous-observation MCTS where darker regions are visited more often. Areas of high reward are generally visited more often by the search as the tree expands.

in 50 simulated trials with the same environments, vehicle, and actions as described in Section IV-A, with the inclusion of 12 block obstacles placed uniformly around the world in known locations (see Fig.4). Boustro. was not used as a baseline because of non-generality of the offline approach to unknown obstacle maps.

As indicated in Table I, PLUMES accumulated significantly more MSS reward than UCB-MCTS and UCB-Myopic, at the 0.05-level. The distribution of reward across the trials is visualized in Fig. 8. Like in the convex-world, the PLUMES has a primary mode between reward 200-250, while the UCB-based planners have a primary mode in the low-performance region (reward <50). There was no significant difference between planners with respect to RMSE or $x^*$ error. The fact that MVI maximized the true MSS reward while achieving statistically indistinguishable error highlights the difference in exploitation efficiency between MVI and UCB-based methods.

The simulation experiments assume that a geometric map is known *a priori*. However in practical applications, like indoor gas leak detection, access to a map may be limited or unavailable. We simulate the scenario in which a nonholonomic car equipped with a laser range-finder must build a map online as it seeks the maximum in a cluttered indoor environment (Fig. 9). We generate a simulated chemical phenomenon from a GP ($l = 0.8$, $\sigma^2 = 100.0$, $\sigma_n^2 = 2.0$ [2%]), and simulate observations at $1\,\mathrm{Hz}$. The action set for the vehicle consists of eleven $1.5\,\mathrm{m}$ Dubins curves projected in front of the vehicle, one straight path behind the vehicle, and a "stay in place" action. Results for five trials are shown in Table I and illustrate that PLUMES accumulates more MSS reward than baselines, indicating robust performance.

These simulation and robot trials demonstrate the utility of PLUMES compared to canonical and state-of-the-art baselines in a diverse set of environments with challenging practical conditions. For high-stakes scientific deployments, the consistent convergence and sampling performance of PLUMES is critical and beneficial.

## V. DISCUSSION AND FUTURE WORK

Online planning methods for robotic maximum seek-and-sample are critical in a variety of contexts, including general environmental monitoring (scientific inquiry, reconnaissance) and disaster response (oil spill, gas leak, radiation). For partially observable environments that can be modelled using a GP, PLUMES is a novel approach for global maximum seek-and-sample that provides several key insights.

This work presents MVI as an empirically suitable alternative to the canonical GP-UCB heuristic in MSS solvers, which is both naturally adaptive and avoids a hand-tuned parameter to balance exploration and exploitation. MVI samples potential global maxima from the robot's full belief state to manage exploration and exploitation. In contrast, heuristic functions like UCB place reward on all high-valued or highly uncertain regions, leading to unnecessary exploration and limiting the time available to exploit knowledge of the true maximum. Ultimately, the MVI heuristic allows PLUMES to collect exploitative samples, while still achieving the same overall level of posterior model accuracy (shown by RMSE) as UCB-based planners. Additionally, continuous-observation MCTS allows PLUMES to search over belief-spaces on continuous functions without discretization or maximum-likelihood assumptions.

One important area of future work for PLUMES is online hyperparameter learning [33], which is important when only one mission is possible and there is insufficient prior knowledge for hyperparameter selection. Another avenue of future work could examine the proprieties of the maxima sampled by MVI, to be used as a heuristic for meta-behavior transitions (e.g., action model switching, dynamic horizon setting) or

mission termination. Finally, the performance of PLUMES in non-convex environments is impacted by the discrete action set encoded. Extending PLUMES to continuous actions spaces, following the approach of Morere et al. [34], would allow flexibility in these environments.

## VI. CONCLUSION

This paper formalizes the maximum-seek-and-sample POMDP and presents PLUMES, an adaptive planning algorithm that employs continuous observation MCTS and maximum-value information reward to perform efficient maximum-seeking in partially observable, continuous environments. PLUMES outperforms canonical coverage and UCB-based state-of-the-art methods with statistical significance in challenging simulated and real-world conditions (e.g. multiple local maxima, unknown obstacles, sensor noise). Maximum seek-and-sample is a critical task in environmental monitoring for which PLUMES, with theoretical convergence guarantees, strong empirical performance, and robustness under real-world conditions, is well-suited.

## ACKNOWLEDGMENT

## REFERENCES

[1] N. Srinivas, A. Krause, S. M. Kakade, and M. W. Seeger, "Information-Theoretic Regret Bounds for Gaussian Process Optimization in the Bandit Setting," *IEEE Trans. Inf. Theory*, vol. 58, no. 5, pp. 3250–3265, may 2012.

[2] J. Binney and G. S. Sukhatme, "Branch and bound for informative path planning," in *2012 IEEE International Conference on Robotics and Automation*. IEEE, 2012, pp. 2147–2154.

[3] Z. W. Lim, D. Hsu, and W. S. Lee, "Adaptive informative path planning in metric spaces," *Int. J. Rob. Res.*, vol. 35, no. 5, pp. 585–598, apr 2016.

[4] A. Arora, P. M. Furlong, R. Fitch, S. Sukkarieh, and T. Fong, "Multi-modal active perception for information gathering in science missions," *International Symposium on Autonomous Robots*, pp. 1–27, 2017.

[5] A. Singh, A. Krause, and W. J. Kaiser, "Nonmyopic adaptive informative path planning for multiple robots." in *IJCAI*, vol. 3, 2009, p. 2.

[6] S. T. Jawaid and S. L. Smith, "Informative path planning as a maximum traveling salesman problem with submodular rewards," *Discret. Appl. Math.*, vol. 186, pp. 112–127, may 2015.

[7] A. Krause, A. Singh, and C. Guestrin, "Near-Optimal Sensor Placements in Gaussian Processes: Theory, Efficient Algorithms and Empirical Studies," *Journal of Machine Learning Research*, vol. 9, pp. 235–284, 2008.

[8] C. K. Ling, K. H. Low, and P. Jaillet, "Gaussian process planning with Lipschitz continuous reward functions: towards unifying Bayesian optimization, active learning, and beyond." in *Association for the Advancement of Artificial Intelligence*, 2016, pp. 1860–1866.

[9] A. Somani, N. Ye, D. Hsu, and W. S. Lee, "DESPOT: Online POMDP planning with regularization," in *Advances in Neural Information Processing Systems*, 2013.

[10] H. Kurniawati and V. Yadav, "An online POMDP solver for uncertainty planning in dynamic environment," in *Robotics Research*. Springer, 2016, pp. 611–629.

[11] D. Silver and J. Veness, "Monte-Carlo planning in large POMDPs," *Advances in Neural Information Processing Systems (NIPS)*, 2010.

[12] Z. N. Sunberg, C. J. Ho, and M. J. Kochenderfer, "The value of inferring the internal state of traffic participants for autonomous freeway driving," in *American Control Conference (ACC), 2017*. IEEE, 2017, pp. 3004–3010.

[13] P. Dallaire, C. Besse, S. Ross, and B. Chaib-draa, "Bayesian reinforcement learning in continuous POMDPs with Gaussian processes," in *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2009, pp. 2604–2609.

[14] R. Marchant, F. Ramos, and S. Sanner, "Sequential Bayesian Optimisation for Spatial-Temporal Monitoring," *UAI*, 2014.

[15] R. Platt Jr, R. Tedrake, L. Kaelbling, and T. Lozano-Perez, "Belief space planning assuming maximum likelihood observations," in *Proceedings of the Robotics: Science and Systems Conference, 6th*, 2010.

[16] A. Couëtoux, J.-B. Hoock, N. Sokolovska, O. Teytaud, and N. Bonnard, "Continuous upper confidence trees," in *International Conference on Learning and Intelligent Optimization*. Springer, 2011, pp. 433–445.

[17] D. Auger, A. Couetoux, and O. Teytaud, "Continuous upper confidence trees with polynomial exploration–consistency," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2013, pp. 194–209.

[18] W. D. Smart and L. P. Kaelbling, "Effective reinforcement learning for mobile robots," in *Proceedings 2002 IEEE International Conference on Robotics and Automation (Cat. No. 02CH37292)*, vol. 4. IEEE, 2002, pp. 3404–3410.

[19] W. Sun, N. Sood, D. Dey, G. Ranade, S. Prakash, and A. Kapoor, "No-regret replanning under uncertainty," in *ICRA*, 2017.

[20] G. Hitz, E. Galceran, M.-È. Garneau, F. Pomerleau, and R. Siegwart, "Adaptive continuous-space informative path planning for online environmental monitoring," *Journal of Field Robotics*, vol. 34, no. 8, pp. 1427–1449, dec 2017.

[21] E. Contal, D. Buffoni, A. Robicquet, and N. Vayatis, "Parallel Gaussian Process Optimization with Upper Confidence Bound and Pure Exploration," in *Machine Learning and Knowedge Discovery in Databases*, 2013, pp. 225–240.

[22] Z. Wang and S. Jegelka, "Max-value entropy search for efficient Bayesian optimization," *arXiv preprint arXiv:1703.01968*, 2017.

[23] L. P. Kaelbling, M. L. Littman, and A. R. Cassandra, "Planning and acting in partially observable stochastic domains," *Artificial intelligence*, vol. 101, no. 1-2, pp. 99–134, 1998.

[24] C. E. Rasmussen and C. K. I. Williams, *Gaussian processes for machine learning*. MIT Press MIT Press, 2004, vol. 14, no. 2.

[25] A. Rahimi and B. Recht, "Random features for large-scale kernel machines," in *Advances in Neural Information Processing Systems*, 2008, pp. 1177–1184.

[26] J. M. Hernández-Lobato, M. W. Hoffman, and Z. Ghahramani, "Predictive entropy search for efficient global optimization of black-box functions," in *Advances in Neural Information Processing Systems*, 2014, pp. 918–926.

[27] H. Choset and P. Pignon, "Coverage path planning: The boustrophedon cellular decomposition," in *Field and service robotics*. Springer, 1998, pp. 203–209.

[28] C. B. Browne, E. Powley, D. Whitehouse, S. M. Lucas, P. I. Cowling, P. Rohlfshagen, S. Tavener, D. Perez, S. Samothrakis, and S. Colton, "A survey of Monte Carlo tree search methods," *IEEE Transactions on Computational Intelligence*, 2012.

[29] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *The annals of mathematical statistics*, pp. 50–60, 1947.

[30] S. Manjanna, N. Kakodkar, M. Meghjani, and G. Dudek, "Efficient terrain driven coral coverage using Gaussian processes for mosaic synthesis," in *13th Conference on Computer and Robot Vision*. IEEE, 2016, pp. 448–455.

[31] A. T. Banaszak and M. P. Lesser, "Effects of solar ultraviolet radiation on coral reef organisms," *Photochemical & Photobiological Sciences*, vol. 8, no. 9, pp. 1276–1294, 2009.

[32] A. Singh, F. Ramos, H. D. Whyte, and W. J. Kaiser, "Modeling and decision making in spatio-temporal processes for environmental surveillance," in *2010 IEEE International Conference on Robotics and Automation*. IEEE, 2010, pp. 5490–5497.

[33] A. Ranganathan, M.-H. Yang, and J. Ho, "Online sparse Gaussian process regression and its applications," *IEEE Transactions on Image Processing*, vol. 20, no. 2, pp. 391–404, 2011.

[34] P. Morere, R. Marchant, and F. Ramos, "Bayesian optimisation for solving continuous state-action-observation POMDPs," *Advances in Neural Information Processing Systems (NIPS)*, 2016.