2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

Planning for Long-Term Monitoring Missions in Time-Varying Environments

Alex Stephens, Bruno Lacerda, Nick Hawes

***Abstract*— Recent years have seen autonomous robots deployed in long-term missions across an ever-increasing breadth of domains. We consider robots deployed over a sequence of finite-horizon missions in the same environment, with the objective of maximising the value from observations of some unknown spatiotemporal process. This work is motivated by applications such as ecological monitoring, in which a robot might be repeatedly deployed in the field over weeks or months with the task of modelling processes of scientific interest. We formalise the problem of long-term monitoring over multiple finite-horizon missions as a Markov decision process with a partially unknown state, and present an online planning approach to address it. Our approach uses a spatiotemporal Gaussian process to model the environment and make predictions about unvisited states, integrating this with a belief-based Monte Carlo tree search algorithm which decides where the robot should go next. We demonstrate the strengths of our framework empirically through a series of experiments using synthetic data as well as real acoustic data from monitoring of bioactivity in coral reefs.**

I. Introduction

As autonomous robots become more capable and reliable, they are increasingly deployed in *long-term* missions for monitoring and information gathering. Robots deployed in such missions must be able to plan effectively in the face of unknown and ever-changing environment dynamics, often with limited and noisy observational data. In this paper, we consider robot missions motivated by applications in ecological monitoring and ocean observation, in which the environment exhibits daily cyclic patterns (such as the influence of weather) combined with slower (e.g. seasonal) variation. In these applications, a mobile robot is deployed in a sequence of missions over a period of weeks or months, with the ability to navigate the environment and take observations before returning to base to recharge or offload data between missions. The mission objective is to maximise the cumulative value of observations taken of some environment feature, about which it has limited prior knowledge. In the ecological setting, features of interest might be plant growth or biodiversity [1], which can be measured by a robot using an on-board sensor.

We formalise the problem of repeated monitoring missions in a partially observable, spatiotemporally-varying environment as a Markov decision process (MDP) with a partially unknown state, and propose an online belief-based planning approach to solve it. At the core of our approach is the robot's model of the unknown environment feature, for which we use a Gaussian process (GP) [2] that captures both spatial and temporal variation. GPs have been used to model continuous environment features in a variety of planning problems [3]–[8], as they provide a flexible and principled predictive framework for settings with sparse observational data. In this work, we assume that the feature of interest varies smoothly with space and time, and focus specifically on features that exhibit periodic behaviour. This reflects our interest in ecological and other outdoor settings which naturally have a daily periodicity, and informs the design of the GP kernel function.

The long-term setting gives rise to significant challenges in scalability, where it quickly becomes infeasible to plan over a full multi-mission horizon even in relatively simple environments. We therefore break the planning problem into individual missions, in which the robot plans its next action online using Monte Carlo tree search (MCTS), forward-simulating the outcomes of different actions up to the end of the current mission using the GP model. Planning only within individual missions runs the risk of encouraging myopic behaviour, so we explicitly reward exploratory actions that can benefit *future* missions. To do so, we plan in a belief-aware manner, balancing exploitation of current knowledge about the environment with exploration guided by GP uncertainty. This requires performing GP belief updates during MCTS, which is computationally costly. Thus, we propose an algorithm that performs GP updates only within the search tree, and estimates the reward for the rollout phase without requiring a belief update at every step. Finally, since we do not assume prior knowledge about the environment feature under observation, the time between missions is used to optimise the GP hyperparameters, allowing the robot to learn a better model of its environment as more observations are made.

The key contributions of this paper are (a) the formulation of long-term, multi-mission robot monitoring of a spatiotemporal process as an MDP with a partially unknown state; (b) a planning approach for this setting using a GP belief over the unknown environment, which balances exploratory information-gathering with reward-seeking behaviour; and (c) a new method for evaluating GP-based rewards during MCTS rollouts for improved computational performance.

II. Related Work

Path planning and decision-making in partially observable environments are topics that have seen significant interest in recent years [3]–[7], [9]–[11]. Due to their high complexity,

All authors are with the Oxford Robotics Institute, University of Oxford, United Kingdom. For correspondence: {stephens, bruno, nickh}@robots.ox.ac.uk.

This work received EPSRC funding via the "From Sensing to Collaboration" programme grant [EP/V000748/1]. A. S. was supported by an Amazon Web Services Lighthouse scholarship and by the AIMS Centre for Doctoral Training. The authors would like to acknowledge the use of the University of Oxford Advanced Research Computing (ARC) facility in carrying out this work (http://dx.doi.org/10.5281/zenodo.22558).

planning problems in partially observable environments are often tackled online with sampling-based methods [12]–[15], which allow for efficient exploration of large state and observation spaces. [15] presents a series of sampling-based approaches to motion planning for robot information gathering, demonstrating asymptotic optimality within a pre-specified cost budget. Other works have adopted a Bayes-adaptive framework [16] to tackle planning problems in the presence of unknown underlying dynamics [6], [17]. In the Bayes-adaptive setting, an agent maintains a history-based belief over the unknown environment dynamics, allowing it to optimally trade-off between exploration and exploitation. Our approach can also be interpreted through the Bayes-adaptive lens. We will discuss this aspect further throughout the paper.

In contrast with the works mentioned so far, our interest is in *long-term*, repeated missions in spatiotemporally varying environments. In our setting, information gathering is an auxiliary goal to serve the long-term reward-maximisation objective – the value of short-term information gathering is in improving our environment model, which can then be exploited later to increase reward overall. There is a significant body of literature addressing persistent monitoring with robots, but these works typically focus on coverage or uncertainty reduction rather than reward gathering [18]–[20].

The problem of predicting the value of *a priori* unknown functions from observational data has been tackled in previous works using GPs, which incorporate uncertainty and provide confidence information that is essential for probabilistic or information-based planning. GPs have been demonstrated as effective tools for modelling a wide variety of natural phenomena, including species distributions [21]–[23], biodiversity [24], dissolved oxygen levels in water [25], and ocean currents [4], [26]. In this work, we use a GP model designed specifically to address the long-term episodic structure of monitoring missions. GPs have also been applied to a variety of robot mission planning problems using MDP models [3]–[8]. Several works perform informative path planning in belief space, using GP beliefs over continuous environments [3], [7], [8]. A GP-based planner is used to address the problem of long-term monitoring of biological activity in [27]. However, this work addresses only scheduling of fixed sensors, whereas our interest is specifically in sensing on a mobile platform, which poses a significantly more complex planning problem.

Belief-space planning using MCTS is computationally costly, and some works have mitigated this using *root sampling*, which reduces the cost of each MCTS iteration by avoiding performing belief updates during them [6], [28]. [6] shows that root sampling from GP beliefs during MCTS produces the same distribution over histories as performing full GP belief updates, thereby providing equivalent planning capability at a fraction of the computational cost. However, in the context of long-term monitoring, it is not feasible to plan over the full mission horizon. Our approach therefore leverages the belief state within the MCTS tree to encourage exploration that might reap benefits *beyond* the planning horizon. We then optimise the efficiency of the planning process by not performing belief updates during the rollout phase, and we demonstrate empirically that the speed-up provided by these approximate rollouts enables better action recommendations under the same planning time budget.

III. Preliminaries

Gaussian processes: A GP is a collection of random variables, any finite number of which are jointly Gaussian distributed [2]. A GP is fully specified by a mean function $m(s)$ and a kernel function $k(s, s')$ parameterised by hyperparameters $\boldsymbol{\theta}$, i.e. $f(s) \sim \mathcal{GP}(m(s), k(s, s'))$. We let $m(s) = 0$ without loss of generality. Given a dataset of $n_{\mathcal{D}}$ noisy observations $\mathcal{D} = \{(s_i, f(s_i) + \epsilon_i)\}_{i=1}^{n_{\mathcal{D}}}$ for states s_i, where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ is Gaussian observation noise, GP regression predicts unknown environment feature values at all inputs s_* based on visited states $s_{\mathcal{D}}$. The hyperparameters can be optimised by maximising their log marginal likelihood given priors $p_0(\boldsymbol{\theta})$ over their values and the observed data $\mathcal{D}$.

MDPs with Unknown Feature Values: We use a finite-horizon MDP with state-based rewards as the basis for our model of the monitoring problem.

A finite-horizon MDP is a tuple $\mathcal{M} = \langle S, s_0, A, T, R, H \rangle$, where S is a set of states, $s_0 \in S$ is the initial state; A is a finite set of actions; $T : S \times A \times S \to [0, 1]$ is the transition function; $R : S \to \mathbb{R}$ is the reward function; and $H \in \mathbb{R}_{>0}$ is the time horizon. We consider actions with deterministic continuous durations, and denote the duration of action $a \in A$ as $dur(a) \in \mathbb{R}_{>0}$. We will address how to incorporate $dur(a)$ into planning when discussing our tree search algorithm. An optimal policy for $\mathcal{M}$ is a function $\pi : S \times [0, H] \to A$ which maximises the sum of rewards up to horizon H.

To clearly separate the known robot dynamics from the unknown environment behaviour, we represent the system as a partially known state MDP (PKSMDP), which we base on the models used in [5], [6]. The definition of a PKSMDP is similar to that of an MDP, but the state space is factored as $S = S_k \times S_e$ where S_k are known features (e.g. the robot's location or resource level) and S_e are unknown features (i.e. the environment process the robot is monitoring). The environment process is defined by an underlying unknown function $f : S_k \to S_e$, which we assume is continuous. Because f is unknown but the value of $f(s_k)$ is uniquely defined, the transition function of the PKSMDP is not a complete transition function, in the sense that it only maps to known state components. Formally, $T : (S_k \times S_e) \times A \times S_k \to [0, 1]$, where $T((s_k, s_e), a, s'_k)$ is the probability of moving to s'_k given that a was executed in state s_k and the value of the environment process was s_e. Note that this formulation allows for transition dynamics that depend on the unknown state component S_e, such as in the case of an underwater robot operating in the presence of unknown currents. Furthermore, the reward function R depends on the full state $S = S_k \times S_e$, and thus considers the unknown state component.

Belief-based tree search with GP observations: We consider planning for a PKSMDP in a setting where f can be observed and modelled as a GP, i.e. the GP is used to maintain a *belief* over the true value of f. This can be viewed as belief-space planning for a Bayes-adaptive MDP, with a GP encoding

the belief over the unknown environment dynamics. Planning in belief space with GPs is computationally expensive, so we use MCTS to generate promising trajectories and select high-reward actions online.

This algorithm is similar to the one presented in [3], and we use their terminology, which refers to planning trees composed of two types of nodes – *belief* nodes, which represent planning states, and *belief-action* nodes, which represent action choices. The root of the search tree is a belief node representing the current state of the mission, which we denote with a timestep index j as $b_j = (s_{k,j}, g_j, s_{e,j})$, where $s_{k,j} \in S_k$ is the known robot state, g_j is a GP posterior over f based on the history of observations *prior* to timestep j, and $s_{e,j} \in S_e$ is an observation sampled from g_j at known state $s_{k,j}$.

Starting with just the root node, trials are performed, either for a fixed number of iterations or until a time budget is exhausted, each consisting of the following steps. In the *selection* step, starting from the root node, an action a is selected according to the PUCT value $\hat{V}(b_j, a) + C\sqrt{\frac{N(b_j)^{e_d}}{N(b_j,a)}}$, where $\hat{V}(b_j, a)$ is the average reward obtained by choosing action a with belief b_j in previous trials; $N(b_j)$ is the number of times that node b_j has been simulated; $N(b_j, a)$ is the number of times that action a has been selected from node b_j; C is a constant that affects the exploration-exploitation balance; and e_d is a depth-dependent parameter [29].

Since the observation space is the co-domain of f, which is continuous, any observation sampled from the GP belief will be unique. Thus, we use *progressive widening* [30] to limit the size of the planning tree, parameterised by depth-dependent parameter α_d [29]. If $\lfloor N(b_j, a)^{\alpha_d} \rfloor = \lfloor (N(b_j, a) - 1)^{\alpha_d} \rfloor$, then the successor state b_{j+1} is randomly sampled from the existing children of node (b_j, a); otherwise, a new state $s_{k,j+1}$ is sampled from $T((s_{k,j}, s_{e,j}), a, .)$ and a new observation $s_{e,j+1}$ is sampled from the GP posterior g_{j+1} at $s_{k,j+1}$. These are used to add a new leaf node to the tree, $b_{j+1} = (s_{k,j+1}, g_{j+1}, s_{e,j+1})$, where the time at the leaf node is given by $t(b_{j+1}) = t(b_j) + dur(a)$.

This process of alternating between action and belief-action nodes continues down the tree until a new (leaf) belief node is added (*expansion*). From this node, we enter the *rollout phase*, where we simulate a sequence of random actions up to the horizon H: that is, until we reach a node b'_j where $t(b'_j) + \min_a dur(a) > H$. This simulated trajectory yields a reward r, which is *backpropagated* to the tree root, updating the average reward and number of queries of each node visited in the selection step. Our method for computing r is described in Section IV-C. Once trials are completed, the action corresponding to the most-visited belief-action child of the root node is selected and executed.

IV. Approach

A. Problem formulation

We consider a mobile robot deployed in a physical environment abstracted as a topological map $G = (V, E)$. Each vertex $v \in V$ represents a location of interest in the environment, i.e. $v = (x, y) \in \mathbb{R}^2$, and edges $e \in E$ represent traversable paths between vertices. The robot can navigate through the environment according to the edges in E. There is an *a priori* unknown spatiotemporal mapping $f : \mathbb{R}^3 \to \mathbb{R}$, where $f(x, y, t)$ is the value of an environment feature of interest at location (x, y) and time t. The objective is to maximise the cumulative reward from observations taken of f across a set of missions that span many days. We model this problem as a Multi-Mission PKSMDP (M^2-PKSMDP), which is defined as a pair $(\mathcal{M}, \mathcal{H})$. The mission PKSMDP is $\mathcal{M} = \langle S, s_0, A, T, R \rangle$, where $\mathcal{S} = \mathcal{S}_k \times \mathcal{S}_e$ with $S_k = \{(x, y, t) \in \mathbb{R}^3 \mid (x, y) \in V \text{ and } t \in \mathbb{R}_{\geq 0}\}$ and $S_e = \mathbb{R}$; $s_0 = (x_0, y_0, 0, f(x_0, y_0, 0))$ where $(x_0, y_0) \in V$ is the initial location; $\mathcal{A} = E$; $\mathcal{T} : (\mathcal{S}_k \times \mathcal{S}_e) \times \mathcal{A} \times \mathcal{S}_k \to [0, 1]$ is the transition function, and; $R : \mathcal{S} \to \mathbb{R}$ is a reward function, which we take to be the value of f, i.e., $R(s_k, f(s_k)) = f(s_k)$. The multi-mission specification $\mathcal{H} = \{(H_i^s, H_i^f)\}_{i=1}^m$ is a list of start and finish times for m consecutive missions.

At the start of each mission i, the robot is deployed at the initial location v_0 at time H_i^s, and explores the environment until the mission time limit H_i^f. Whenever the robot arrives at a state $s_{k,j}$, it receives a noisy observation z_j of the unknown function f subject to measurement noise $\epsilon \sim \mathcal{N}(0, \sigma_z^2)$, with $z_j = f(s_{k,j}) + \epsilon_j$. We refer to the sequence of known states visited and observations received during mission i as a *trajectory* $\sigma^i = s_{k,1}^i z_1^i s_{k,2}^i z_2^i \ldots s_{k,n_i}^i z_{n_i}^i$, where $s_{k,1}^i$ is the known component of the initial state of the i-th mission, i.e., $s_{k,1}^i = (s_0, t_0, H_i^s)$. The full multi-mission trajectory is denoted as $\boldsymbol{\sigma} = \{\sigma^1, \ldots, \sigma^m\}$.

For the remainder of this paper, based on our motivating setting of ecological monitoring, we consider the missions to be *daily*, such that the multi-mission problem described by $(\mathcal{M}, \mathcal{H})$ spans $|\mathcal{H}| = m$ consecutive days. We also assume for notational simplicity that f models a single environment feature only. However, the approach can easily generalise to multiple unknown environment features using a multi-output GP [31], by redefining the reward to be a scalar function of the multiple features.

The monitoring objective is to find a policy π that maximises the expected cumulative reward across all missions:

$$V_\pi(s_0) = E_\pi \left[\sum_{\sigma^i \in \boldsymbol{\sigma}} \sum_{s_j^i \in \sigma^i} R\left(s_j^i\right) \right]. \tag{1}$$

The core challenges of tackling this problem are in (a) effective modelling of the spatiotemporal process f with sparse observational data; (b) planning over multiple missions, balancing exploratory behaviour that may benefit future missions with exploitation of knowledge from past ones; and (c) scaling to large numbers of missions for truly long-term monitoring scenarios.

B. Spatiotemporal modelling with GPs

Temporal variation amplifies the challenge of building accurate predictive models with limited observational data. By accounting for the structure of the problem at hand in the design of our GP kernel function, we can reduce the amount of data required to accurately model the unknown process [32].

Our design encapsulates two key structural assumptions about the domains of interest: first, that the process varies smoothly in both space and time, and second, that it consists of an underlying cyclic component.

In the following, let $\mathbf{x} = (x, y) \in \mathbb{R}^2$, $\mathbf{x}' = (x', y') \in \mathbb{R}^2$, $s = (\mathbf{x}, t) \in \mathbb{R}^3$ and $s' = (\mathbf{x}', t') \in \mathbb{R}^3$, Our kernel function can be expressed as a product of a spatial and temporal component $k(s, s') = k_{\mathbf{x}}(\mathbf{x}, \mathbf{x}')k_t(t, t')$. The spatial component is a radial basis function (RBF):

$$k_{\mathbf{x}}(\mathbf{x}, \mathbf{x}') = \sigma_{\mathbf{x}}^2 \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\ell_{\mathbf{x}}^2}\right), \tag{2}$$

and the temporal component is a product of a periodic and RBF kernel:

$$k_t(t, t') = \sigma_t^2 \exp\left(-\frac{2\sin^2(\pi\,|t - t'|\,/p)}{\ell_p^2}\right) \exp\left(-\frac{(t - t')^2}{2\ell_q^2}\right). \tag{3}$$

Here, $\sigma_{\mathbf{x}}$ and σ_t are the spatial and temporal variances respectively; $\ell_{\mathbf{x}}$ is the spatial lengthscale; p is the period of the temporal kernel; ℓ_p is the lengthscale of the periodic component; and ℓ_q is the lengthscale of the slowly-varying component. On its own, a periodic kernel assumes *purely* periodic behaviour; in combination with the RBF kernel, some periodicity is preserved, but points that are further apart in time are treated as less correlated, thereby allowing for slow variation between cycles with $\ell_q > p$. We hold the value of the period $p = 24\,\mathrm{h}$ fixed, capturing an assumption that the environment will feature daily variation. The remaining hyperparameters mentioned above are optimised between episodes by maximising the log marginal likelihood for the model given the current history of observations.

C. Evaluating trajectory rewards

The state-dependent reward in our mission objective (Equation 1) depends on the unknown function f, over which the GP maintains a belief. We use the GP belief at step j, denoted g_j, to compute an upper confidence bound (UCB) [33] reward for belief state $b_j = (s_{k,j}, g_j, s_{e,j})$, given by:

$$\tilde{R}_{\mathrm{UCB}}(s_{k,j}, g_j) = \mu_j(s_{k,j}) + \kappa\sigma_j(s_{k,j}), \tag{4}$$

where μ_j and σ_j are respectively the mean and standard deviation of g_j at state $s_{k,j}$, and κ is a constant that affects the exploration-exploitation balance. Note that the known state $s_{k,j}$ has both spatial and temporal components, so this reward is also spatiotemporal. The UCB reward adds an exploration bonus to states with greater uncertainty, which is essential in our multi-mission setting since we are only able to plan over a single-mission horizon. Exploratory actions that seem unfavourable within the current mission may still contribute to increased reward in future missions, and the UCB reward encourages taking such actions.

Thus, the reward accumulated over a simulated trajectory $\sigma = s_{k,1}z_1 s_{k,2}z_2 \ldots s_{k,n}z_n$ using the UCB reward is:

$$R^+(\sigma) = \sum_{j=1}^{n} \tilde{R}_{\mathrm{UCB}}(s_{k,j}, g_j), \tag{5}$$

where g_j is the GP posterior given dataset $z_1 z_2 \ldots z_j$. However, evaluating each $\tilde{R}_{\mathrm{UCB}}(s_j, g_j)$ requires sampling an observation from the previous GP belief g_{j-1}, and updating a copy of that belief. This is computationally costly, and reduces the capacity to perform sufficient MCTS trials to accurately estimate the value of each action at the root node.

To mitigate this, we estimate reward accumulated during the rollout phase using the GP belief g_ℓ from the *leaf* node b_ℓ; i.e., we do not perform belief updates during this phase. Formally, for a trajectory $\sigma = s_{k,1}z_1 s_{k,2}z_2 \ldots s_{k,n}z_n$ with $1 \leq \ell \leq n$ being the index where the search moves to the rollout phase, we define the estimated cumulative reward as:

$$\tilde{R}^+(\sigma) = \sum_{j=0}^{\ell} \tilde{R}_{\mathrm{UCB}}(s_{k,j}, g_j) + \sum_{j=\ell+1}^{n} \tilde{R}_{\mathrm{UCB}}(s_{k,j}, g_\ell), \tag{6}$$

where g_j is the GP posterior given dataset $z_1 z_2 \ldots z_j$.

This approach removes the need to create and sample from many copies of the GP during the rollout phase, instead allowing the reward accumulated during this rollout phase to be estimated through a single GP query. The in-tree component of each MCTS iteration is still quite computationally intensive. However, since the rewards depend explicitly on belief uncertainty, this cannot be easily circumvented, and we show in Section V that we are still able to obtain good performance within reasonable computation times.

D. Planning algorithm

Our planning approach is summarised in Algorithm 1. Each mission i begins with the robot at location $\mathbf{x}_0 = (x_0, y_0)$, at the mission start time $t = H_i^s$. Actions are selected online (Line 10) using belief-based MCTS as described above, which plans to the *current* mission horizon using a GP model based on the history of observations up to that point. The selected action is then executed (Line 11), generating a new observation which is added to the dataset and used to update the GP model (Line 12). GP hyperparameters are optimised offline during the robot's down time between missions (Line 15), improving the model over the course of the mission.

V. Experiments

A. Experimental domains

We demonstrate the performance of our algorithm on domains based on acoustic monitoring of coral reef activity. Studies of these environments have found that sound levels at low frequencies correlate with density and diversity of coral reef fish species, which vary periodically over both daily and seasonal timescales [27], [34]. The mission goal is to focus observations of the reef on areas where bioactivity is highest, using acoustic sensing data to inform the GP model.

Algorithm 1 MULTI-MISSION MONITORING

Input: M^2-PKSMDP $(\mathcal{M}, \mathcal{H})$ where $\mathcal{M} = \langle S, s_0, A, T, R\rangle$ and $\mathcal{H} = \{(H_i^s, H_i^f)\}_{i=1}^m$, GP hyperparameter priors $\boldsymbol{\theta}_0$, kernel $k(s, s')$
Output: GP posterior over environment feature

$i \leftarrow 1$ (current mission)
$\mathcal{D} \leftarrow \emptyset$
$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}_0$
Initialise GP model $G_{\mathcal{D},\boldsymbol{\theta}}$ with kernel k, hyperparameters $\boldsymbol{\theta}$, dataset $\mathcal{D}$
while $i \leq |H|$ **do**
  $\mathbf{x} \leftarrow \mathbf{x}_0$
  $t \leftarrow H_i^s$
  $s \leftarrow (\mathbf{x}, t)$
  while $t \leq H_i^f$ **do**
    $a \leftarrow$ MCTSACTIONSELECTION$(s, G_{\mathcal{D},\boldsymbol{\theta}})$
    $s', o' \leftarrow$ EXECUTEACTION(a)
    $\mathcal{D} \leftarrow \mathcal{D} \cup \{(s', o')\}$
    $s \leftarrow$ new state s'
  end while
  $\boldsymbol{\theta} \leftarrow$ OPTIMISEHYPERPARAMETERS$(G_{\mathcal{D},\boldsymbol{\theta}})$
  $i \leftarrow i + 1$
end while

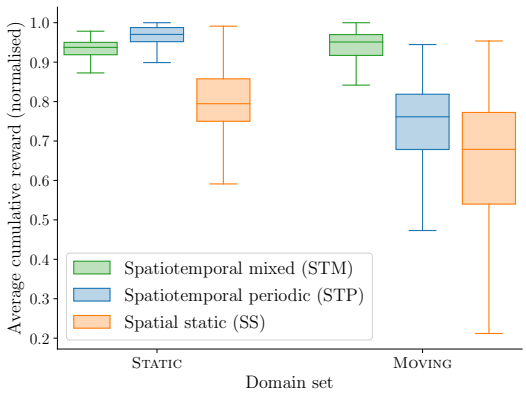


Fig. 1: Average reward over 10 STATIC and 10 MOVING domains, using planners with different GP kernels. Rewards are normalised per-domain to the highest reward obtained on that domain across all complete runs. 10 runs per domain.

Synthetic domains: We have 10 STATIC domains and 10 MOVING domains, which capture to different behaviours of the unknown environment feature f, each defined over an 8×8 grid map. In the STATIC domains, f consists of a set of randomly-placed Gaussian point sources at *fixed* locations, whose amplitudes oscillate periodically over 24 hours; each source is spatially static, but still varies over time. In these domains, f is purely periodic, capturing daily variation only. In the MOVING domains, f consists of 1-2 *moving* Gaussian point sources, which oscillate in amplitude in the same manner as the static sources but also drift slowly across the map, emulating gradual seasonal variation. The design of these domains reflect patterns seen in many ecological phenomena, which often display similar combinations of daily cycles and longer-term variation. In each case, the multi-mission duration is 20 days, and the robot is operational for 8 hours each day, performing 5 actions per hour.

Coral reef acoustic domain: The SESOKO domain uses acoustic spectrogram data captured at three sites in a coral reef near the island of Sesoko, Japan [35]. Figure 2a shows a portion of the raw data. The signal power at each site is computed as a normalised root-mean-square across the 400-600 Hz range, with some smoothing to remove high-frequency noise. Each of these three signals is then used to modulate the amplitude of a Gaussian point source placed within a 10×10 grid environment. We also add a small drift velocity to one of the sources, to simulate possible long-term movement which could not be explicitly captured by the static acoustic sensors. These multi-missions span 30 days, with the robot active for 20 hours per day, performing 3 actions per hour.

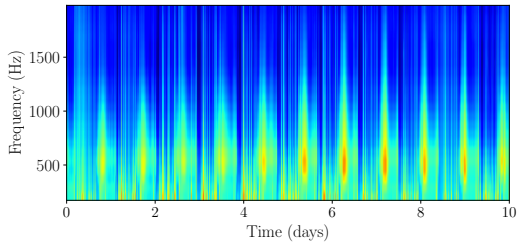
B. Results

Experiment 1: Planning with a spatiotemporal GP kernel: This experiment set investigates whether our spatiotemporal GP kernel is able to produce improved planning behaviour in practice, compared with simpler kernels. We run 10 multi-missions in each of the 10 STATIC and MOVING domains, testing three different GP kernels. The first is our full spatiotemporal kernel $k(\mathbf{s}, \mathbf{s'})$ from Equations 2–3. Second, we remove the gradual temporal variation component from Equation 3, so the kernel is purely periodic in time. Finally, we remove the temporal kernel entirely, using a purely spatial RBF kernel as an additional baseline. We refer to these kernel configurations respectively as spatiotemporal mixed (STM), spatiotemporal periodic (STP), and spatial static (SS).

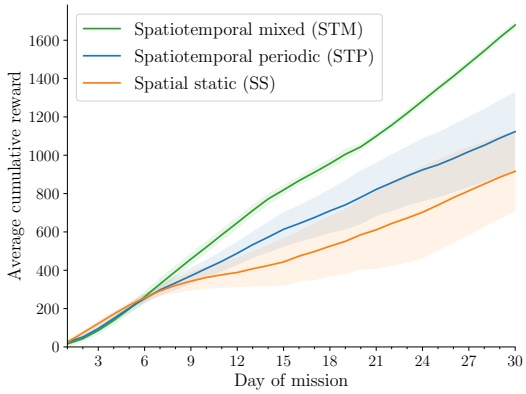
We see in Figure 1 that in the STATIC domains, STP performs best, since the unknown function in these domains is completely periodic and thus can be captured accurately by a kernel with a periodic temporal component. The STM kernel performs well, although its performance is reduced slightly by the presence of the non-periodic temporal component. This is as expected, since the additional term in k_t (Equation 3) reduces the weight of older observations in GP inference – in a purely periodic environment, this is a slight disadvantage, resulting in predictions being made using less data. The SS kernel performs worst, since it is unable to capture temporal variation; it can only identify the region of the map with the highest *average* reward across all missions.

In the MOVING domains, where bioactivity changes gradually between days, Figure 1 shows that the performance of both the STP and SS kernels are weaker than the STM kernel. In these domains, it is no longer possible to simply identify the best location to visit at particular times of day, since these locations change over time. However, the STM kernel is able to learn and exploit an accurate model after a few days, and continues to perform well across all subsequent missions.

Finally, in the SESOKO domain, Figure 2b shows that the planner with the STM kernel achieves the highest reward, with the SS and STP kernel performance degrading considerably in the latter half of the mission. The STM kernel captures the underlying periodicity in the acoustic data, despite significant noise and artifacts in the raw data. This validates our GP kernel design, showing that the combination of periodic and RBF kernels enables effective modelling and planning using limited data in a real-world robotic monitoring scenario.

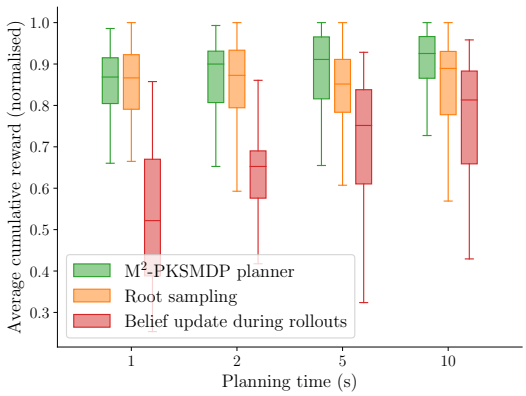


(a) Frequency spectrogram data from Site C, north of Sesoko island.

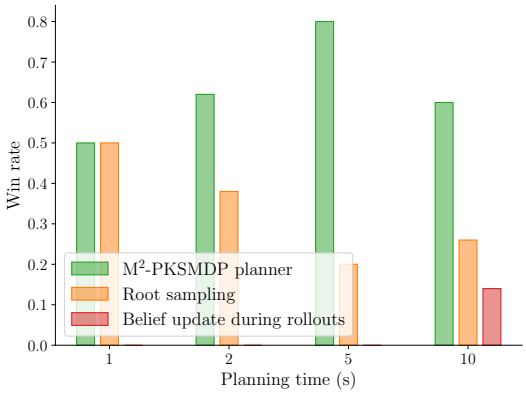


(b) Average cumulative reward across 10 30-day missions in the SESOKO environment; shaded regions indicate standard deviation.

Fig. 2: Experimental data and results from the coral reef acoustic monitoring dataset from Sesoko, Japan [35].



(a) Average cumulative reward (normalised per-domain).



(b) Win rate, indicating the proportion of runs on which each algorithm obtained the highest reward

Fig. 3: Planner performance under fixed per-action planning time. Each configuration was run 5 times on each of the 10 MOVING domains.

Experiment 2: Belief-based MCTS across multiple missions: We now investigate how the core components of our M^2-PKSMDP planner contribute to performance. Our planner uses a belief-based MCTS algorithm which rewards exploration of uncertain states using a belief-dependent reward. This requires sampling from and updating a GP model each time a new node is added to the search tree. During the rollout phase, however, trajectory rewards are evaluated in a single batch *without* updating the GP, allowing for fast evaluation across many actions. We therefore compare our approach to two baselines which remove these elements. The first uses root sampling similar to [6], where instead of updating the GP during MCTS, each observation is sampled from the belief at the root node, and no exploration bonus is used. This approach is shown to converge to the reward-maximising policy and significantly reduces the time taken for each trial. However, it does not allow for the explicit rewarding of exploration provided by the belief-dependent rewards. Thus, the policies from [6] greedily optimise for each mission, without accounting for the benefits that exploration in one mission might have on performance in later missions. The second baseline modifies the rollout phase to update the GP at each step, which results in a more accurate trajectory reward estimate at the expense of more computation.

Figure 3a compares the rewards obtained by the three algorithms in the MOVING domains when given different amounts of planning time. We see that performing GP belief updates during the rollout phase of MCTS does not improve overall performance, indicating that computation time is better spent simply performing *more* trials. However, given more computation time, the performance of this method slowly improves, as it represents a more accurate value estimation process. In particular, it is able to win a few of the test runs when given 10 seconds of planning time per decision step. However, this already adds significant planning overhead, with 10 seconds per decision step being infeasible in many domains. Furthermore, our planner with 1 second planning time still outperforms planning with GP updates during rollouts with 10 seconds planning time. The root sampling baseline performs similarly to the M^2-PKSMDP planner when planning time per action is limited to 1 s. However, as we increase the planning time, we see that the M^2-PKSMDP planner outperforms the root sampling method, obtaining higher average cumulative reward and a win rate (Figure 3b) of at least 60% in all cases.

From these results, we conclude that integrating a belief-based exploration reward can improve overall performance on long-term missions. However, to achieve this improvement, updating of the GP belief should be done only when traversing the search tree, while the rollout phase should only consider a fixed belief from the leaf node it started from.

VI. Conclusion

We have presented the M^2-PKSMDP, a formalism for long-term, repeated robot monitoring missions in the presence of *a priori* unknown environment features. Motivated by applications in ecological monitoring, we employ a GP kernel designed to capture spatiotemporally-varying processes, with a periodic component to model these features. This GP is then integrated into a continuous-observation MCTS algorithm that plans online, balancing exploitation of observation from past missions with exploration to benefit future ones. In future work, we plan to extend the planning framework to consider longer multi-missions in more complex environments. Interesting avenues could include learning causal models of multiple features, and incorporating influences from external factors such as temperature readings and weather forecasts.

References

[1] J. Jackson, C. S. Lawson, C. Adelmant, E. Huhtala, P. Fernandes, R. Hodgson, H. King, L. Williamson, K. Maseyk, N. Hawes *et al.*, "Short-range multispectral imaging is an inexpensive, fast, and accurate approach to estimate biodiversity in a temperate calcareous grassland," *Ecology and Evolution*, vol. 12, no. 12, p. e9623, 2022.

[2] C. K. Williams and C. E. Rasmussen, *Gaussian processes for machine learning*. MIT press Cambridge, MA, 2006, vol. 2.

[3] G. Flaspohler, V. Preston, A. P. Michel, Y. Girdhar, and N. Roy, "Information-guided robotic maximum seek-and-sample in partially observable continuous environments," *IEEE Robotics and Automation Letters*, vol. 4, no. 4, pp. 3782–3789, 2019.

[4] P. Duckworth, B. Lacerda, and N. Hawes, "Time-bounded mission planning in time-varying domains with semi-mdps and gaussian processes," in *Conference on Robot Learning*. PMLR, 2021, pp. 1654–1668.

[5] M. Budd, B. Lacerda, P. Duckworth, A. West, B. Lennox, and N. Hawes, "Markov decision processes with unknown state feature values for safe exploration using gaussian processes," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 7344–7350.

[6] M. Budd, P. Duckworth, N. Hawes, and B. Lacerda, "Bayesian reinforcement learning for single-episode missions in partially unknown environments," in *Conference on Robot Learning*. PMLR, 2023, pp. 1189–1198.

[7] P. Morere, R. Marchant, and F. Ramos, "Sequential bayesian optimization as a pomdp for environment monitoring with uavs," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 6381–6388.

[8] R. Marchant, F. Ramos, S. Sanner *et al.*, "Sequential bayesian optimisation for spatial-temporal monitoring." in *UAI*. Citeseer, 2014, pp. 553–562.

[9] C. Costen, M. Rigter, B. Lacerda, and N. Hawes, "Shared autonomy systems with stochastic operator models." International Joint Conferences on Artificial Intelligence Organization, 2022.

[10] S.-K. Kim, A. Bouman, G. Salhotra, D. D. Fan, K. Otsu, J. Burdick, and A.-a. Agha-mohammadi, "Plgrim: Hierarchical value learning for large-scale exploration in unknown environments," in *Proceedings of the International Conference on Automated Planning and Scheduling*, vol. 31, 2021, pp. 652–662.

[11] J. Capitan, M. T. Spaan, L. Merino, and A. Ollero, "Decentralized multi-robot cooperation with auctioned pomdps," *The International Journal of Robotics Research*, vol. 32, no. 6, pp. 650–671, 2013.

[12] H. Kurniawati and V. Yadav, "An online pomdp solver for uncertainty planning in dynamic environment," in *Robotics Research: The 16th International Symposium ISRR*. Springer, 2016, pp. 611–629.

[13] D. Silver and J. Veness, "Monte-carlo planning in large pomdps," *Advances in neural information processing systems*, vol. 23, 2010.

[14] A. Somani, N. Ye, D. Hsu, and W. S. Lee, "Despot: Online pomdp planning with regularization," *Advances in neural information processing systems*, vol. 26, 2013.

[15] G. A. Hollinger and G. S. Sukhatme, "Sampling-based robotic information gathering algorithms," *The International Journal of Robotics Research*, vol. 33, no. 9, pp. 1271–1287, 2014.

[16] M. O. Duff, *Optimal Learning: Computational procedures for Bayes-adaptive Markov decision processes*. University of Massachusetts Amherst, 2002.

[17] C. Costen, M. Rigter, B. Lacerda, and N. Hawes, "Planning with hidden parameter polynomial mdps," in *AAAI Conference on Artificial Intelligence*, 2022.

[18] J. Binney, A. Krause, and G. S. Sukhatme, "Optimizing waypoints for monitoring spatiotemporal phenomena," *The International Journal of Robotics Research*, vol. 32, no. 8, pp. 873–888, 2013.

[19] X. Lan and M. Schwager, "Rapidly exploring random cycles: Persistent estimation of spatiotemporal fields with multiple sensing robots," *IEEE Transactions on Robotics*, vol. 32, no. 5, pp. 1230–1244, 2016.

[20] K.-C. Ma, Z. Ma, L. Liu, and G. S. Sukhatme, "Multi-robot informative and adaptive planning for persistent environmental monitoring," in *Distributed Autonomous Robotic Systems: The 13th International Symposium*. Springer, 2018, pp. 285–298.

[21] N. Golding and B. V. Purse, "Fast and flexible bayesian species distribution modelling using gaussian processes," *Methods in Ecology and Evolution*, vol. 7, no. 5, pp. 598–608, 2016.

[22] M. Ingram, D. Vukcevic, and N. Golding, "Multi-output gaussian processes for species distribution modelling," *Methods in ecology and evolution*, vol. 11, no. 12, pp. 1587–1598, 2020.

[23] W. J. Wright, K. M. Irvine, T. J. Rodhouse, and A. R. Litt, "Spatial gaussian processes improve multi-species occupancy models when range boundaries are uncertain and nonoverlapping," *Ecology and Evolution*, vol. 11, no. 13, pp. 8516–8527, 2021.

[24] M. V. Talluto, K. Mokany, L. J. Pollock, and W. Thuiller, "Multifaceted biodiversity modelling at macroecological scales using gaussian processes," *Diversity and Distributions*, vol. 24, pp. 1492–1502, 10 2018.

[25] L. Bottarelli, M. Bicego, J. Blum, and A. Farinelli, "Orienteering-based informative path planning for environmental monitoring," *Engineering Applications of Artificial Intelligence*, vol. 77, pp. 46–58, 2019.

[26] D. Sarkar, M. A. Osborne, and T. A. Adcock, "Spatiotemporal prediction of tidal currents using gaussian processes," *Journal of Geophysical Research: Oceans*, vol. 124, no. 4, pp. 2697–2715, 2019.

[27] S. McCammon, N. Aoki, T. A. Mooney, and Y. Girdhar, "Adaptive online sampling of periodic processes with application to coral reef acoustic abundance monitoring," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 11 671–11 678.

[28] A. Guez, D. Silver, and P. Dayan, "Scalable and efficient bayes-adaptive reinforcement learning based on monte-carlo tree search," *Journal of Artificial Intelligence Research*, vol. 48, pp. 841–883, 2013.

[29] D. Auger, A. Couetoux, and O. Teytaud, "Continuous upper confidence trees with polynomial exploration–consistency," in *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23-27, 2013, Proceedings, Part I 13*. Springer, 2013, pp. 194–209.

[30] A. Couëtoux, J.-B. Hoock, N. Sokolovska, O. Teytaud, and N. Bonnard, "Continuous upper confidence trees," in *Learning and Intelligent Optimization: 5th International Conference, LION 5, Rome, Italy, January 17-21, 2011. Selected Papers 5*. Springer, 2011, pp. 433–445.

[31] M. A. Osborne, S. J. Roberts, A. Rogers, S. D. Ramchurn, and N. R. Jennings, "Towards real-time information processing of sensor network data using computationally efficient multi-output gaussian processes," in *2008 international conference on information processing in sensor networks (IPSN 2008)*. IEEE, 2008, pp. 109–120.

[32] N. D. Goodman, T. D. Ullman, and J. B. Tenenbaum, "Learning a theory of causality." *Psychological review*, vol. 118, no. 1, p. 110, 2011.

[33] N. Srinivas, A. Krause, S. M. Kakade, and M. W. Seeger, "Information-theoretic regret bounds for gaussian process optimization in the bandit setting," *IEEE transactions on information theory*, vol. 58, no. 5, pp. 3250–3265, 2012.

[34] E. Staaterman, M. B. Ogburn, A. H. Altieri, S. J. Brandl, R. Whippo, J. Seemann, M. Goodison, and J. E. Duffy, "Bioacoustic measurements complement visual biodiversity surveys: preliminary evidence from four shallow marine habitats," *Marine Ecology Progress Series*, vol. 575, pp. 207–215, 2017.

[35] T.-H. Lin, T. Akamatsu, F. Sinniger, and S. Harii, "Exploring coral reef biodiversity via underwater soundscapes," *Biological Conservation*, vol. 253, p. 108901, 2021.