Supplementary Material S2 - Bayesian design methods for improving the effectiveness of ecosystem monitoring

Pubudu Thilan AWL[1,2,3,8], Erin Peterson[1,2,7], Patricia Menndez[4], Julian Caley[1,2], Christopher Drovandi[1,2,3], Camille Mellin[5,6], James McGree[1,2,3*]

*Corresponding author(s). E-mail(s): james.mcgree@qut.edu.au;
Contributing authors: pubudu@maths.ruh.ac.lk;
erin@peterson-consulting.com; patricia.menendez@unimelb.edu.au;
julian.caley@gmail.com; c.drovandi@qut.edu.au;
camille.mellin@adelaide.edu.au;

Supplementary Material S2: Further details about adaptive designs

Section S1: Approximate the expected utility

The approach to evaluate our approximation to the expected utility is outlined in Algorithm S1. To apply this algorithm, firstly a design and a number of quantities need to be initialised/defined (line 1). Then, for a large value of T (≥ 100), parameter values are simulated from the prior (i.e. the prior for design; line 3), time-varying disturbances are simulated from an assumed distribution (line 4) (describe in the next section) and data are simulated from the likelihood (given the simulated values of the random effect, parameter, and time-varying disturbances; line 5). Based on the simulated data, a posterior distribution is found. Here, for computational efficiency, this is approximated by the Laplace approximation (McGree et al 2016; Overstall et al 2018). Employing this approach requires the use of the full data likelihood (i.e. not conditional on random effects). In our case, this is not available in closed-form, so needs to be approximated. Again, Monte Carlo integration is used for this purpose (line 6). Based on this posterior distribution, the KLD utility is evaluated (line 7)

(Eq. S1), and this value is stored (line 8). Once this process is completed T times, the average of the utility values is taken as the approximation to the expected utility.

Algorithm S1: Approximate the expected utility

---

1. Initialize $\boldsymbol{d} = (\boldsymbol{d}_1, \boldsymbol{d}_2, \cdots, \boldsymbol{d}_i, \cdots, \boldsymbol{d}_\nu)$, where $\nu \leq 27$

   $\boldsymbol{\theta} = (\beta_0, \boldsymbol{\beta}_d, \boldsymbol{\beta}_z, \beta_k, \psi, \sigma_r^2, \phi), \boldsymbol{\kappa}_{p_0}$, T and C

2. For $t = 1$ to T do

3. Simulate $\boldsymbol{\theta}^{(t)} \sim p(\boldsymbol{\theta} \mid \boldsymbol{y}_{p_0}, \boldsymbol{z}_{p_0}, \boldsymbol{d}_{p_0})$

4. Simulate $\boldsymbol{z}^{(t)} \sim p(\boldsymbol{z}|\boldsymbol{d}, \boldsymbol{\kappa}_{p_0}, \boldsymbol{z}_{p_0}, \boldsymbol{d}_{p_0})$, $\boldsymbol{z} = z_{ik}, i = 1, \cdots, \nu, k = 9$

5. Simulate $\boldsymbol{y}^{(t)} \sim p(\boldsymbol{y}|\boldsymbol{\theta}^{(t)}, \boldsymbol{z}^{(t)}, \boldsymbol{r}^{(c)}, \boldsymbol{d}, \boldsymbol{y}_{p_0}, \boldsymbol{z}_{p_0}, \boldsymbol{d}_{p_0})$, $\boldsymbol{r}^{(c)} = (r_i, i = 1, \cdots, \nu)$

6. Estimate $p(\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{d}, \boldsymbol{y}_{p_0}, \boldsymbol{z}_{p_0}, \boldsymbol{d}_{p_0})$ via Laplace approximation

   with an approximated likelihood $p(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{d}, \boldsymbol{y}_{p_0}, \boldsymbol{z}_{p_0}, \boldsymbol{d}_{p_0})$

   $\approx \frac{1}{C} \sum_{c=1}^{C} p(\boldsymbol{y} \mid \boldsymbol{z}^{(t)}, \boldsymbol{\theta}^{(t)}, \boldsymbol{r}^{(c)}, \boldsymbol{d}, \boldsymbol{y}_{p_0}, \boldsymbol{z}_{p_0}, \boldsymbol{d}_{p_0})$

7. Evaluate KLD utility $u(\boldsymbol{d}, \boldsymbol{z}^{(t)}, \boldsymbol{y}^{(t)}|\boldsymbol{y}_{p_0}, \boldsymbol{z}_{p_0}, \boldsymbol{d}_{p_0})$

8. Store $u(\boldsymbol{d}, \boldsymbol{z}^{(t)}, \boldsymbol{y}^{(t)}|\boldsymbol{y}_{p_0}, \boldsymbol{z}_{p_0}, \boldsymbol{d}_{p_0})$

9. End for

10. Output $E[u(\boldsymbol{d}, \boldsymbol{z}, \boldsymbol{y}|\boldsymbol{y}_{p_0}, \boldsymbol{z}_{p_0}, \boldsymbol{d}_{p_0})] \approx \frac{1}{T} \sum_{t=1}^{T} u(\boldsymbol{d}, \boldsymbol{z}^{(t)}, \boldsymbol{y}^{(t)}|\boldsymbol{y}_{p_0}, \boldsymbol{z}_{p_0}, \boldsymbol{d}_{p_0})$

Section S2: Distribution of time-varying covariates

To evaluate the approximation to the expected utility function, time-varying disturbances $\boldsymbol{z}$ need to be simulated for each site ($i = 1, \cdots, \nu$) for the next sampling year ($k = 9$) (Algorithm S1, line 4). Thus, distributions $p(\boldsymbol{z}|\boldsymbol{d}, \boldsymbol{\kappa}_{p_0}, \boldsymbol{z}_{p_0}, \boldsymbol{d}_{p_0})$ for these covariates need to be specified where $\boldsymbol{\kappa}_{p_0}$ is an unknown site-specific parameter. To estimate this parameter, moment matching was used (Vose 2008) based on site-level data $\boldsymbol{z}_{p_0}$. Following this, in the case of categorical covariates (i.e. bleaching and cyclone impacts; Table 1), we estimated the proportion of observed occurrences of each disturbance for each site. Then, the outcome of disturbance present or not was modelled, assuming it follows a Bernoulli distribution. To develop a distribution for CoTS, we first determined the proportion of sites where no CoTS were recorded (Zeileis et al 2008). Then, the outcome of CoTS being zero or not was modelled via a Bernoulli distribution. When CoTS was not equal to zero, a Log-normal distribution was estimated and used to simulate future values. Thus, to simulate CoTS data, we first generated a random number q_i ($i = 1, \cdots, \nu$), between 0 and 1, and if $q_i < p_i$ (i.e. proportion of CoTS=0 at the site), we set CoTS = 0, otherwise, we generated CoTS data from the fitted

Log-normal distribution. In doing so, distributions of time-varying covariates were determined, and the corresponding estimated parameters at the site level are shown in Table S1. In evaluating designs, these parameter values were used to simulate three time-varying covariate impacts for the set of sites within a considered design. In the table, the first column includes the 27 sites surveyed in the Whitsunday region and the estimated parameters related to time-varying covariates at each site are provided in the remaining columns. These distributions were then used to simulate time-varying disturbances to approximate the expected utility, as shown in Algorithm S1.

Section S3: Utility approximation

To evaluate the approximation to the expected utility, see Eq. (5), T posterior distributions need to be approximated using the Laplace approximation. As the resulting posterior distributions are Multivariate Normal (also the prior), KLD between the prior and posterior distribution can be evaluated as follows:

$$u(\boldsymbol{d}, \boldsymbol{y}|\boldsymbol{d}_{p_0}, \boldsymbol{y}_{p_0}) = \frac{1}{2}\Bigg(\mathrm{tr}((\boldsymbol{\Sigma}^*_{m_1})^{-1}\boldsymbol{\Sigma}^*_{m_0}) + (\boldsymbol{\theta}^*_{m_1} - \boldsymbol{\theta}^*_{m_0})^T \times (\boldsymbol{\Sigma}^*_{m_1})^{-1} \times (\boldsymbol{\theta}^*_{m_1} - \boldsymbol{\theta}^*_{m_0}) - q_m + \ln\left(\frac{\det\boldsymbol{\Sigma}^*_{m_1}}{\det\boldsymbol{\Sigma}^*_{m_0}}\right)\Bigg), \tag{S1}$$

where $\boldsymbol{\theta}^*_{m_0}$, $\boldsymbol{\theta}^*_{m_1}$ and $\boldsymbol{\Sigma}^*_{m_0}$, $\boldsymbol{\Sigma}^*_{m_1}$ are the means and variance-covariance matrices of the prior and posterior distributions, respectively and q_m is the dimension of the parameter space.

Section S4: Efficiency evaluation

To compare designs $\boldsymbol{d}^K$ based on methods from Kang et al (2016) with our optimal designs $\boldsymbol{d}^*$, the following design efficiency was evaluated:

$$\mathrm{Eff}_\omega = \frac{E[u(\boldsymbol{d}^K_\omega, \boldsymbol{z}, \boldsymbol{y}|\boldsymbol{y}_{p_0}, \boldsymbol{z}_{p_0}, \boldsymbol{d}_{p_0})]}{E[u(\boldsymbol{d}^*_\omega, \boldsymbol{z}, \boldsymbol{y}|\boldsymbol{y}_{p_0}, \boldsymbol{z}_{p_0}, \boldsymbol{d}_{p_0})]}, \;\; \omega = 1, \cdots, 20, \tag{S2}$$

where $E[u(\boldsymbol{d}^K_\omega, \boldsymbol{z}, \boldsymbol{y}|\boldsymbol{y}_{p_0}, \boldsymbol{z}_{p_0}, \boldsymbol{d}_{p_0})]$ and $E[u(\boldsymbol{d}^*_\omega, \boldsymbol{z}, \boldsymbol{y}|\boldsymbol{y}_{p_0}, \boldsymbol{z}_{p_0}, \boldsymbol{d}_{p_0})]$ are the ωth evaluations of the approximate expected utilities evaluated using Eq. (5) for the optimal design under the linear modelling approach from Kang et al (2016) and the optimal design under our spatial model, respectively. As the expected utility evaluation is stochastic (this evaluation was based on $T \geq 100$ number of simulation), each design was evaluated 20 times. Then, the mean efficiency was evaluated as an average of $\mathrm{Eff}_\omega = 1, 2, \cdots, 20$.

Similarly, to compare our optimal designs $\boldsymbol{d}^*$ with the LTMP design $\boldsymbol{d}^L$, design efficiency was again used. In the case of exploring optimal designs under a reduced survey approach compared to the LTMP (i.e. under the "Impacts of reduced survey"), the efficiency was based on the expected utility of the design $\boldsymbol{d}^L$ as follows:

Table S1 The estimated parameters for the distributions of time-varying disturbances at each site in the Whitsunday region

Site number	Bleaching proportion	Cyclone proportion	CoTS proportion	log CoTS mean	log CoTS standard deviation
1	0.12	0.25	0.62	-4.44	2.80
2	0.12	0.25	0.62	-4.60	2.99
3	0.12	0.25	0.62	-4.83	3.27
4	0.12	0.12	0.62	-3.84	2.88
5	0.12	0.12	0.62	-3.82	2.86
6	0.12	0.12	0.62	-3.80	2.85
7	0.12	0.12	0.62	-4.31	2.85
8	0.12	0.12	0.62	-4.29	2.81
9	0.12	0.12	0.62	-4.30	2.83
10	0.12	0.37	0.75	-5.45	3.90
11	0.13	0.38	0.77	-5.28	3.73
12	0.12	0.37	0.75	-5.28	3.73
13	0.12	0.37	0.75	-9.70	1.41
14	0.12	0.37	0.75	-9.62	1.40
15	0.12	0.37	0.75	-9.39	1.40
16	0.12	0.50	0.75	-6.90	2.43
17	0.13	0.49	0.77	-6.65	2.27
18	0.12	0.50	0.75	-6.43	2.12
19	0.12	0.25	0.75	-8.45	1.47
20	0.12	0.25	0.75	-8.14	1.47
21	0.12	0.25	0.75	-7.96	1.47
22	0.12	0.37	0.75	-7.17	2.18
23	0.12	0.37	0.75	-6.86	2.05
24	0.12	0.37	0.75	-6.62	1.96
25	0.12	0.37	0.75	-8.26	2.28
26	0.12	0.37	0.75	-8.26	2.28
27	0.13	0.36	0.77	-8.26	2.28

$$\text{Eff}_\omega = \frac{E[u(\boldsymbol{d}^*_\omega, \boldsymbol{z}, \boldsymbol{y} | \boldsymbol{y}_{p_0}, \boldsymbol{z}_{p_0}, \boldsymbol{d}_{p_0})]}{E[u(\boldsymbol{d}^L_\omega, \boldsymbol{z}, \boldsymbol{y} | \boldsymbol{y}_{p_0}, \boldsymbol{z}_{p_0}, \boldsymbol{d}_{p_0})]}, \;\; \omega = 1, \cdots, 20, \tag{S3}$$

where $E[u(\boldsymbol{d}^*_\omega, \boldsymbol{z}, \boldsymbol{y}|\boldsymbol{y}_{p_0}, \boldsymbol{z}_{p_0}, \boldsymbol{d}_{p_0})]$ and $E[u(\boldsymbol{d}^L_\omega, \boldsymbol{z}, \boldsymbol{y}|\boldsymbol{y}_{p_0}, \boldsymbol{z}_{p_0}, \boldsymbol{d}_{p_0})]$ are the ωth evaluations of approximate expected utilities evaluated using Eq. (5) for $\boldsymbol{d}^*$ and $\boldsymbol{d}^L$, respectively. As previously, the mean efficiency was evaluated as an average of $\text{Eff}_\omega = 1, 2, \cdots, 20$. Furthermore, under the "Impacts of different disturbance scenarios", the efficiency was evaluated (and reported) with respect to our optimal designs $\boldsymbol{d}^*$, i.e. considering the inverse of Eq. (S3) (as there will be no survey reduction here).

Section S5: Impacts of different disturbance conditions

Two scenarios were examined to determine designs that vary with the time-varying environmental disturbances. In Scenario 1, we simulated environmental disturbances to match the historical disturbance patterns in the Whitsunday region. The results of these simulations are presented and discussed in detail within the manuscript. In Scenario 2, we determined optimal designs subject to CoTS disturbance under four survey schemes as described in Section 3.3. The sites selected with the optimal design under these four schemes are shown in Fig. S1 and S2. The number of samples that each optimal design collected from CoTS affected/unaffected sites under these four schemes is shown using a dot plot in Fig. S3. It is evident from these figures that the optimal designs collect data from some CoTS affected and unaffected sites, presumably to estimate impact. The mean efficiencies of the design $\boldsymbol{d}^L$ with reference to these optimal designs were 47 (0.47)%, 48 (0.57)%, 50 (0.59)%, and 51 (0.69)%, respectively. These relatively low efficiencies suggest $\boldsymbol{d}^*$ is substantially outperforming $\boldsymbol{d}^L$ in addressing our monitoring objective.

Under the hypothetical disturbance scheme (i), the optimal design does not collect data from each affected site (e.g. Site 5) (Fig. S1 (a)-(c) and Fig. S3 (a)). It seems that collecting data across all affected sites is not needed to reasonably estimate impact. Moreover, if the optimal design has already collected data from a CoTS affected site on a nearby reef, it appears that the optimal design may not sample another affected site but rather an unaffected site to capture some other variations. This is evident in the inshore where the optimal design has collected more data from the CoTS affected site on Hayman reef, but instead from the CoTS affected site, more data have been collected from an unaffected site on nearby Langford-bird reef.

Similarly, under the scheme (ii), where we considered all the inshore sites as affected sites, the optimal design does not sample all the affected sites (Fig. S1 (d) and Fig. S3 (b)). Instead, fewer sites are sampled in the affected area compared to the current LTMP design where samples would be collected from all the sites. We found a similar pattern of sampled reefs in the middle-shelf (scheme (iii)) and the outer-shelf (scheme (iv)) habitats (Fig. S2 and Fig. S3 (c) and (d)). These results indicate that one can use adaptive design with relatively less survey effort to relatively efficiently quantify the impact of environmental disturbances such as those considered in this paper.

Section S6: Qualitative comparison of the optimal designs

In this section, we compare the designs found under our Beta regression model and the linear model proposed by Kang et al (2016) (Fig. S4). One of the main extensions of our model is the inclusion of the spatial random effect that accounts for the fact

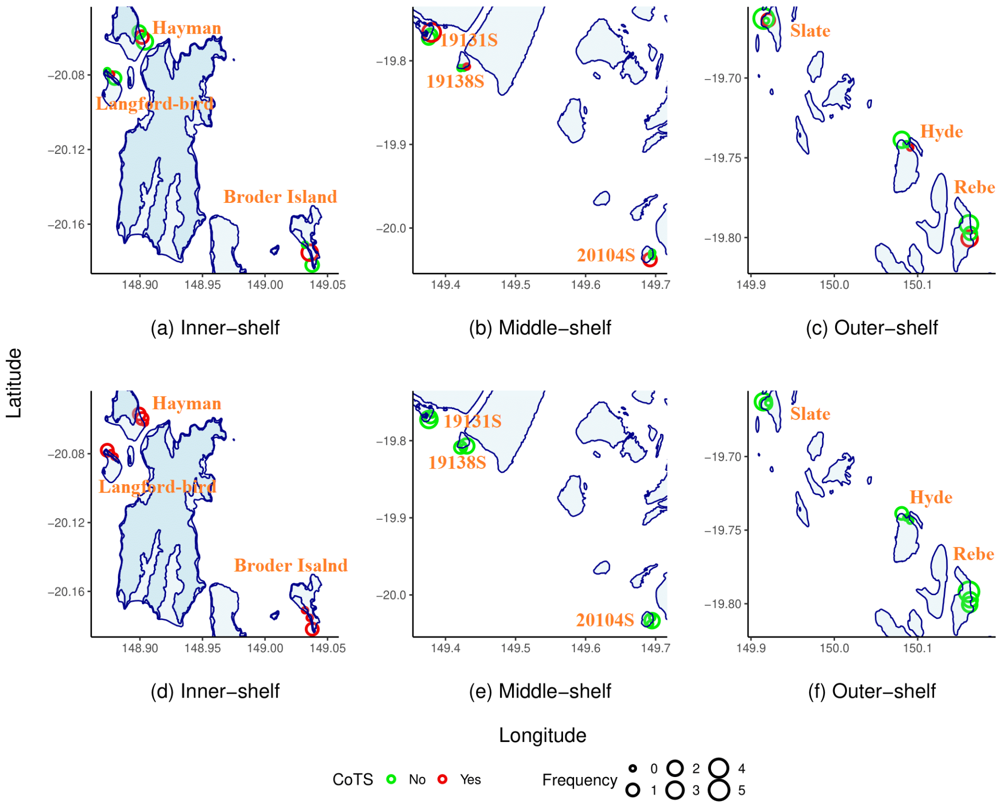

Fig. S1 The optimal designs under the scenarios of CoTS disturbance for one selected site on each reef ((a), (b), and (c)) and for all the inshore-shelf sites ((d), (e), and (f)). This figure was produced using the GRB shape file having a nominal scale of 1:250,000 (Great Barrier Reef Marine Park Authority 2014) using *R* package *ggplot2* (Wickham 2011). In each panel, the Whitsunday region of the Great Barrier Reef is divided into three parts based on habitat as Inner- (left), Middle- (middle), and Outer-shelf (right). Green and red circles represent "unaffected" and "affected" sites, respectively. Frequency represents the number of observations to be collected from "unaffected" and "affected" sites. A small amount of jitter was added for visualisation purposes

that observations collected in space may not be independent. To explore this, we can inspect the estimated range parameter reported in Table 5. On a standardised distance scale, the estimated range is 0.33. After un-standardising this value, the estimated range is approximately equal to 12.76km. This implies that coral cover at reefs/sites separated by distances less than 12.76km are spatially correlated, whereas reefs/sites further than 12.76km apart are not. With this in mind, we compared designs selected based on the linear and Beta regression models. As can be seen, based on our model, Hayman reef is selected as one of the least informative reefs (Fig. 4). This reef is within 12.76km of Langford-bird reef implying that information about Hayman reef can be obtained by sampling at this reef. In contrast, Broder Island was selected as the least

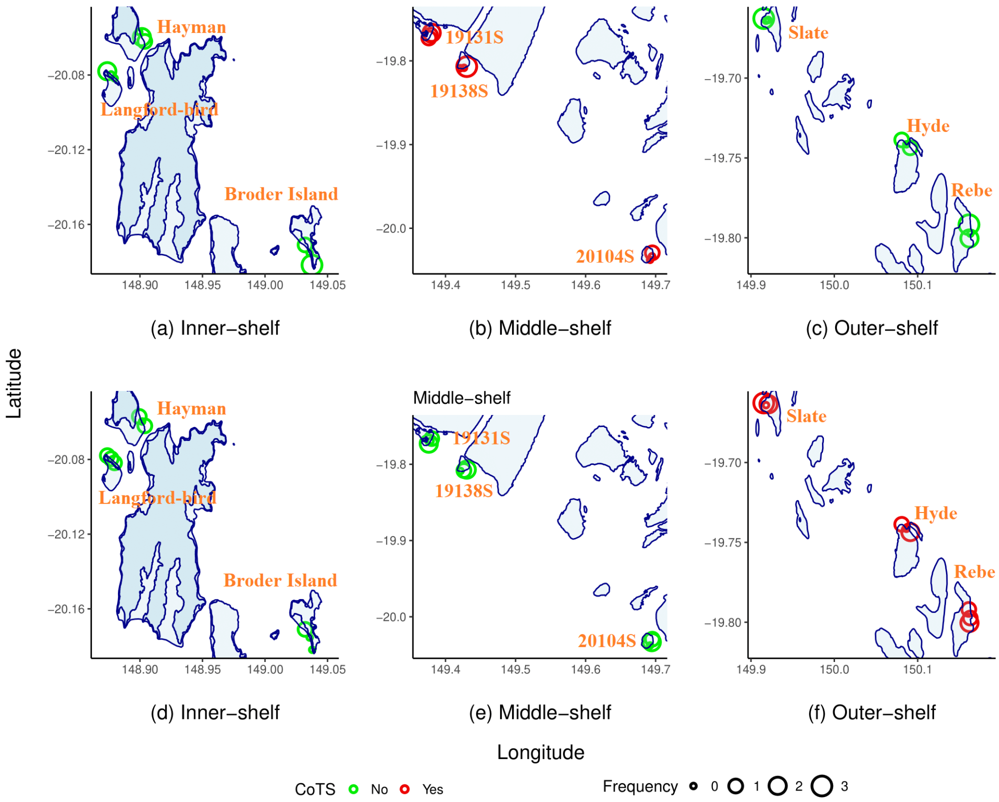

Fig. S2 The optimal designs under the scenarios of CoTS disturbance for all the middle-shelf sites ((a), (b), and (c)) and for all the outer-shelf sites ((d), (e), and (f)). This figure was produced using the GRB shape file having a nominal scale of 1:250,000 (Great Barrier Reef Marine Park Authority 2014) using *R* package *ggplot2* (Wickham 2011). In each panel, the Whitsunday region of the Great Barrier Reef is divided into three parts based on habitat as Inner- (left), Middle- (middle), and Outer-shelf (right). Green and red circles represents “unaffected” and “affected” sites, respectively. Frequency represent the number of observations to be collected from “unaffected” and “affected” sites. A small amount of jitter was added for visualisation purposes

informative reef based on the linear model, which is relatively isolated in space. Thus, the designs found based on our modelling approach appear to leverage the information that can be obtained due to significant spatial variability in coral cover leading to the exclusion of reefs within a close vicinity to others.

Next, we compared the performance of our optimal designs against random surveys in two selected examples. First, when disturbance patterns are similar to those historically observed (Scenario 1), we compared the performance of our optimal design against random selection as well as the corresponding optimal design found based on methods from Kang et al (2016). From Fig. S5 (a), it can be seen that our optimal design outperforms the random designs as well as the designs based on the approach

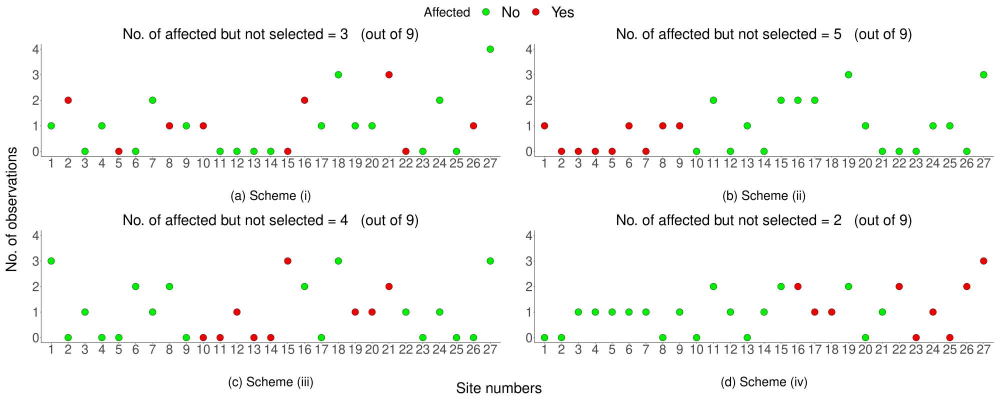

Fig. S3 Number of observations to be collected from CoTS affected and unaffected sites under the four schemes considered in this study. These four schemes include (a) one site from each reef affected, (b) all sites in the inshore reefs affected, (c) all sites in the middle-shelf reefs affected, and (d) all sites in the outer-shelf reefs affected

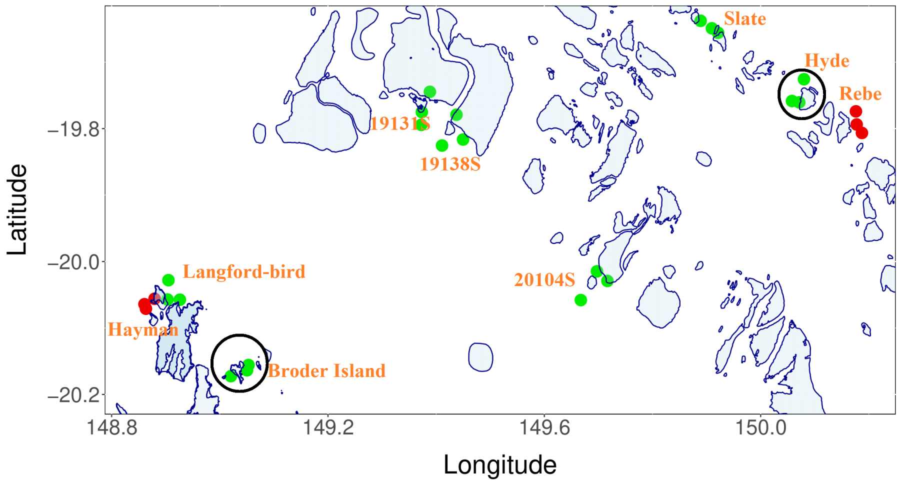

Fig. S4 Visualisation of spatial locations of the two least informative reefs (sites) in the Whitsunday region of the Great Barrier Reef under spatial model and Kang et al. (2016). Reefs (sites) removed based on the Beta regression model are shown in red while those based on the linear regression model are circled in black. A small amount of jitter was added for visualisation purposes

of Kang et al (2016). Indeed, the optimal design maintains an average efficiency of around 250% compared to the LTMP design. Second, when one site from each reef is disturbed by a CoTS outbreak (scheme (i) under Scenario 2), the optimal design outperforms the random designs and yields an average efficiency of greater than 200%, see Fig. S5 (b). These results confirm that our optimal designs provide highly informative data for addressing our monitoring objective compared to the LTMP, Kang et al (2016), and random designs, under two selected examples.

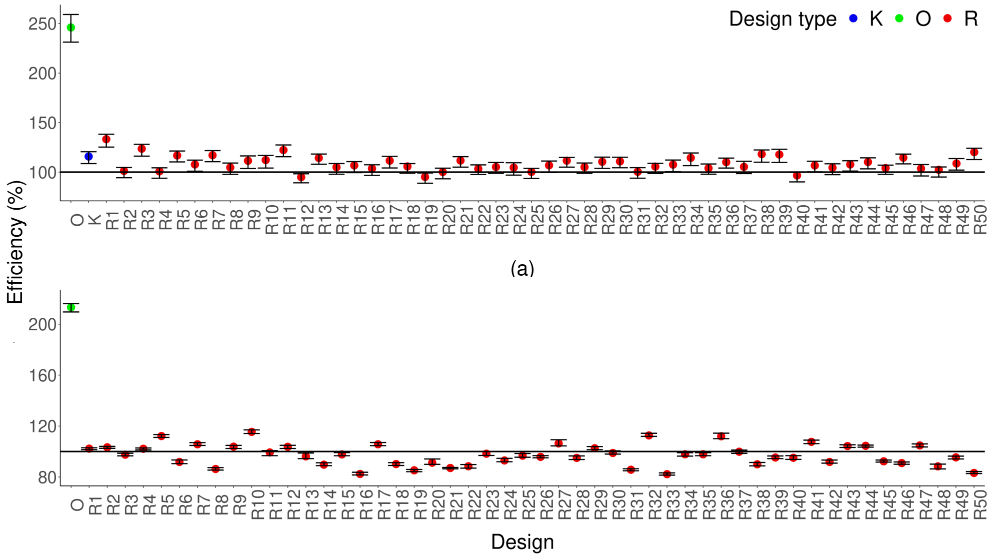

Fig. S5 The efficiency comparison with random designs. (a) The efficiency comparison of our optimal design with Kang et al (2016) and random designs when disturbance patterns match historical disturbance pattern in the region and (b) the efficiency comparison of our optimal design with Kang et al (2016) and random designs under the scheme of one site from each reef affected from CoTS in the Whitsunday region. Here, O denotes the optimal design, K denotes the design based on methods from Kang et al (2016) and R denotes a random design. The efficiencies were calculated with respect to the Long-Term Monitoring program of the Great Barrier Reef. This comparison was performed with 100 random designs. For the visualisation purpose, only the first 50 random designs were considered. The black horizontal line represents the efficiency value of 100

References

Great Barrier Reef Marine Park Authority (2014) Great Barrier Reef (GBR) Features (Reef boundaries, QLD Mainland, Islands, Cays, Rocks and Dry Reefs) (GBRMPA). Available from https://geoportal.gbrmpa.gov.au/. Accessed 07 September 2023

Kang SY, McGree JM, Drovandi CC, et al (2016) Bayesian adaptive design: improving the effectiveness of monitoring of the Great Barrier Reef. Ecological Applications 26(8):2637–2648. https://doi.org/https://doi.org/10.1002/eap.1409

McGree JM, Drovandi CC, White G, et al (2016) A pseudo-marginal sequential Monte Carlo algorithm for random effects models in Bayesian sequential design. Statistics and Computing 26(5):1121–1136. https://doi.org/https://doi.org/10.1007/s11222-015-9596-z

Overstall AM, McGree JM, Drovandi CC (2018) An approach for finding fully bayesian optimal designs using normal-based approximations to loss functions. Statistics and Computing 28(2):343–358. https://doi.org/https://doi.org/10.1007/s11222-017-9734-x

Vose D (2008) Risk analysis: a quantitative guide. John Wiley & Sons

Wickham H (2011) ggplot2. Wiley Interdisciplinary Reviews: Computational Statistics 3:180–185

Zeileis A, Kleiber C, Jackman S (2008) Regression models for count data in R. Journal of Statistical Software 27(8):1–25