Supplementary Material S1 - Bayesian design methods for improving the effectiveness of ecosystem monitoring

Pubudu Thilan AWL[1,2,3,8], Erin Peterson[1,2,7], Patricia Menéndez[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 S1 – Further details about the statistical model

Section S1: Posterior estimation

In modelling our historical data, we adopted a vague prior to ensure that all inferences are prodominately data-driven. The specific prior selected for each component of $\boldsymbol{\theta}$ is shown in Table S1.

Table S1 Prior distribution for modelling the LTMP data

Parameters	**Prior distributions**
β_0	$N(0, 100^2)$
Each component of $\boldsymbol{\beta_d}$	$N(0, 100^2)$
Each component of $\boldsymbol{\beta_z}$	$N(0, 100^2)$
β_k	$N(0, 100^2)$
$\log(1/\psi)$	$N(0, 2)$
$\log\ \sigma_r^2$	$N(-2, 1)$
$\log\ \phi$	$N(-2, 1)$

Given a design $\boldsymbol{d}_{p_0}$ with response data $\boldsymbol{y}_{p_0}$, the posterior distribution can be defined as $p(\boldsymbol{\theta}|\boldsymbol{y}_{p_0}, \boldsymbol{z}_{p_0}, \boldsymbol{d}_{p_0}) \propto p(\boldsymbol{y}_{p_0}|\boldsymbol{\theta}, \boldsymbol{z}_{p_0}, \boldsymbol{d}_{p_0})p(\boldsymbol{\theta})$, where $p(\boldsymbol{\theta})$ is the prior and $p(\boldsymbol{y}_{p_0}|\boldsymbol{\theta}, \boldsymbol{z}_{p_0}, \boldsymbol{d}_{p_0})$ is the likelihood. As random effects are included in the model, the likelihood can be defined as follows:

$$p(\boldsymbol{y}_{p_0}|\boldsymbol{\theta}, \boldsymbol{z}_{p_0}, \boldsymbol{d}_{p_0}) = \int_{\boldsymbol{r}} p(\boldsymbol{y}_{p_0}|\beta_0, \boldsymbol{\beta_d}, \boldsymbol{\beta_z}, \beta_t,\ \psi, \boldsymbol{r}, \boldsymbol{z}_{p_0}, \boldsymbol{d}_{p_0}) \\ p(\boldsymbol{r}|\, \sigma_r^2,\ \phi)d\boldsymbol{r}, \tag{S1}$$

where $p(\boldsymbol{y}_{p_0}|\beta_0, \boldsymbol{\beta_d}, \boldsymbol{\beta_z}, \beta_t,\ \psi,\ \boldsymbol{r}, \boldsymbol{z}_{p_0}, \boldsymbol{d}_{p_0})$ is the likelihood conditional on the random effects and $p(\boldsymbol{r}|\, \sigma_r^2,\ \phi)$ is the distribution of the random effects. In general, the above integral cannot be solved analytically, so Monte Carlo integration was used to approximate the likelihood as follows:

$$p(\boldsymbol{y}_{p_0}|\boldsymbol{\theta}, \boldsymbol{z}_{p_0}, \boldsymbol{d}_{p_0}) \approx \frac{1}{C}\sum_{c=1}^{C} p(\boldsymbol{y}_{p_0}|\beta_0, \boldsymbol{\beta_d}, \boldsymbol{\beta_z}, \beta_t,\ \psi,\ \boldsymbol{r}^{(c)},\ \boldsymbol{z}_{p_0}, \boldsymbol{d}_{p_0}), \tag{S2}$$

where $\boldsymbol{r}^{(c)} \sim p(\boldsymbol{r}|\, \sigma_r^2,\ \phi)$. For the application of the Laplace approximation, the likelihood with the random effects integrated out is needed (i.e. Eq. S1). Thus, the approximation obtained using Eq. S2 can be used to approximate the posterior distribution via the Laplace approximation, and has the following form:

$$\boldsymbol{\theta}|\boldsymbol{y}_{p_0}, \boldsymbol{z}_{p_0}, \boldsymbol{d}_{p_0} \sim \mathrm{MVN}(\boldsymbol{\theta}^*, \boldsymbol{\Sigma}^*), \tag{S3}$$

where $\boldsymbol{\theta}^*$ denotes the mode of the posterior distribution and $\boldsymbol{\Sigma}^* = [-H(\boldsymbol{\theta}^*)]^{-1}$ denotes the variance-covariance matrix being the inverse of the negative Hessian matrix evaluated at $\boldsymbol{\theta}^*$.

Section S2: Validate the fitted statistical model

To validate the fitted model and compare it with the model from Kang et al (2016), a posterior predictive check was used where we compared the 95% posterior predictive intervals for the two models (Fig. S1). As can be seen, both models appear to capture the average behaviour of coral cover. However, there are differences when comparing variability. In particular, approximately 1% and 4% of the observed data fall outside these intervals for Kang et al (2016) and spatial Beta regression models, respectively. This suggests that the spatial Beta regression model is preferred over the Kang et al (2016) model as we expect 5% of the data to lie outside these intervals.

Additionally, we assessed the goodness-of-fit of the spatial Beta regression model by considering posterior predictive data at each site level (Fig. S2-S4). It seems from the figures that our model captures the trend and variability in the data at the site level well.

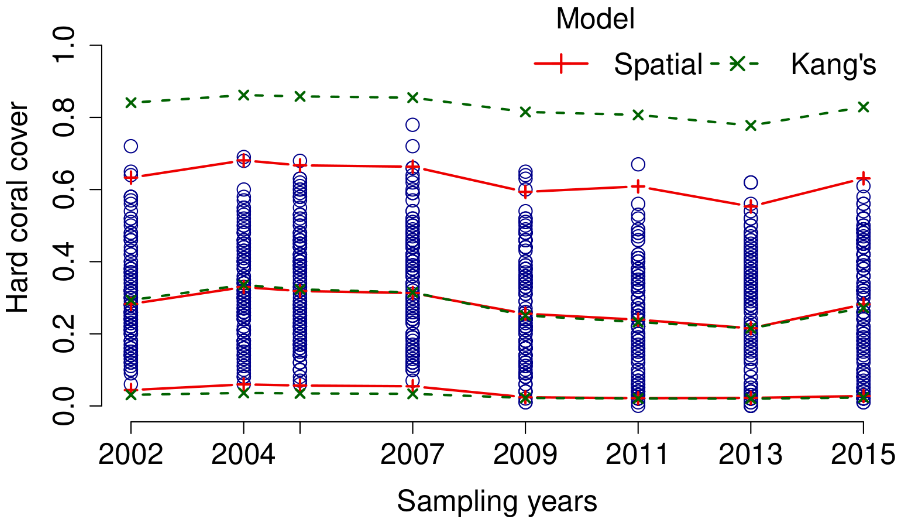

Fig. S1 Posterior predictive comparison between the spatial and Kang et al (2016) models. This includes the median and the lower and upper bounds of 95% posterior predictive intervals for both models

To determine the impact of prior choices on model parameters, a sensitivity analysis was conducted (Saltelli et al 2000). For this purpose, a range of possible priors were considered one at-a-time for each parameter. The results showed that the posterior summaries were relatively insensitive to such prior choice. As an additional check of model fit, block cross-validation was implemented for predicting "unseen" data. This block cross-validation approach is more appropriate than random cross-validation when data include structures, for example, in space and time (Roberts et al 2017). Accordingly, the data were split into two groups as: "seen" data, all data except data for the last survey year (i.e. 2015), and "unseen" data, last year of survey data. Then, the posterior predictive distribution was used to assess how the model can predict the "unseen" data. The posterior predictive distribution for the "unseen" data is shown in Fig. S5. As shown, the "unseen" data generally reside within the predictive intervals. This provides evidence that this model is appropriate for these data.

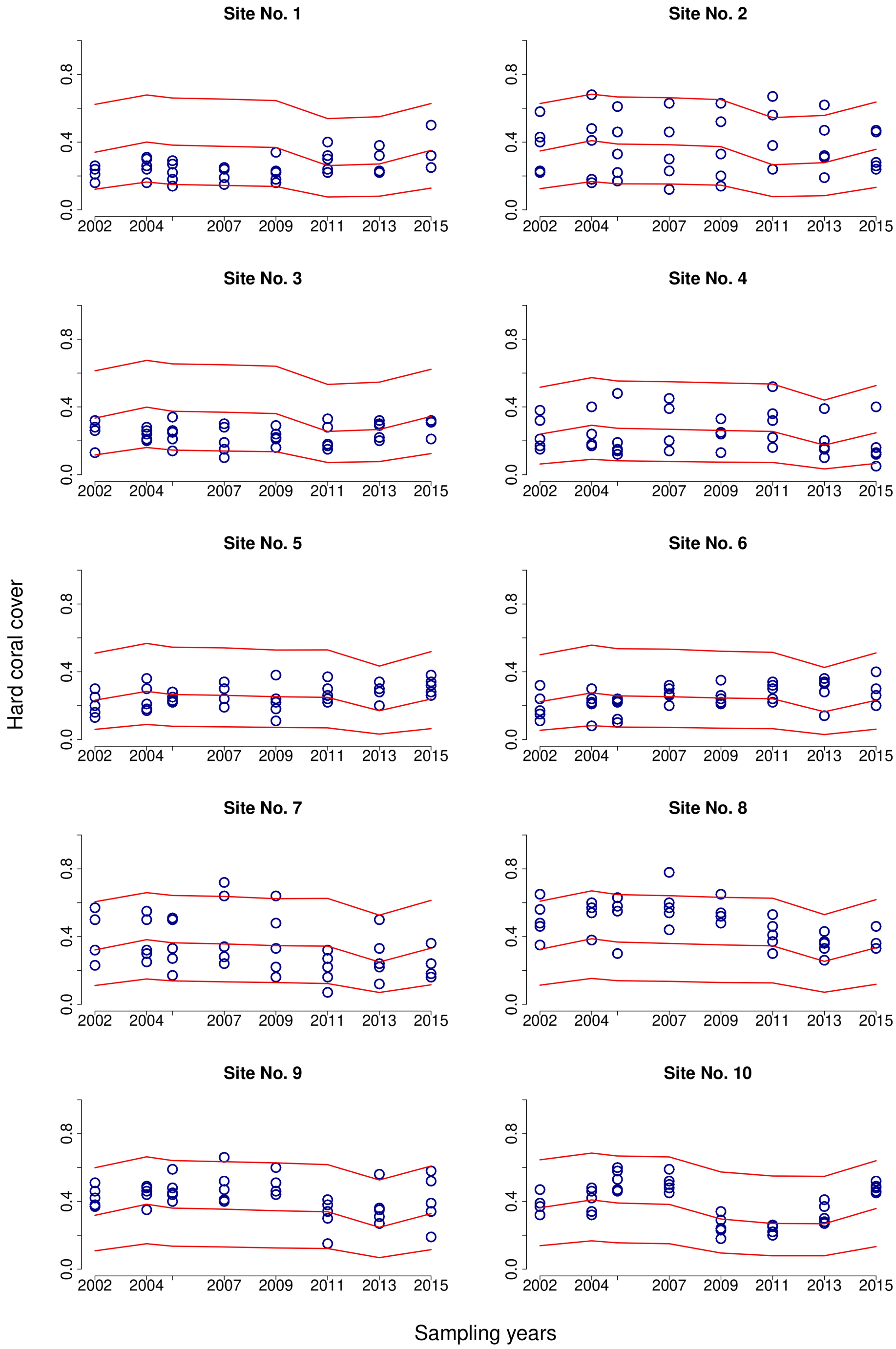

Fig. S2 Posterior predictive by sites for sites 1-10 in the Whitsunday region of the Great Barrier Reef. Red lines represent the median and the lower and upper bounds of 95% credible intervals for the prediction

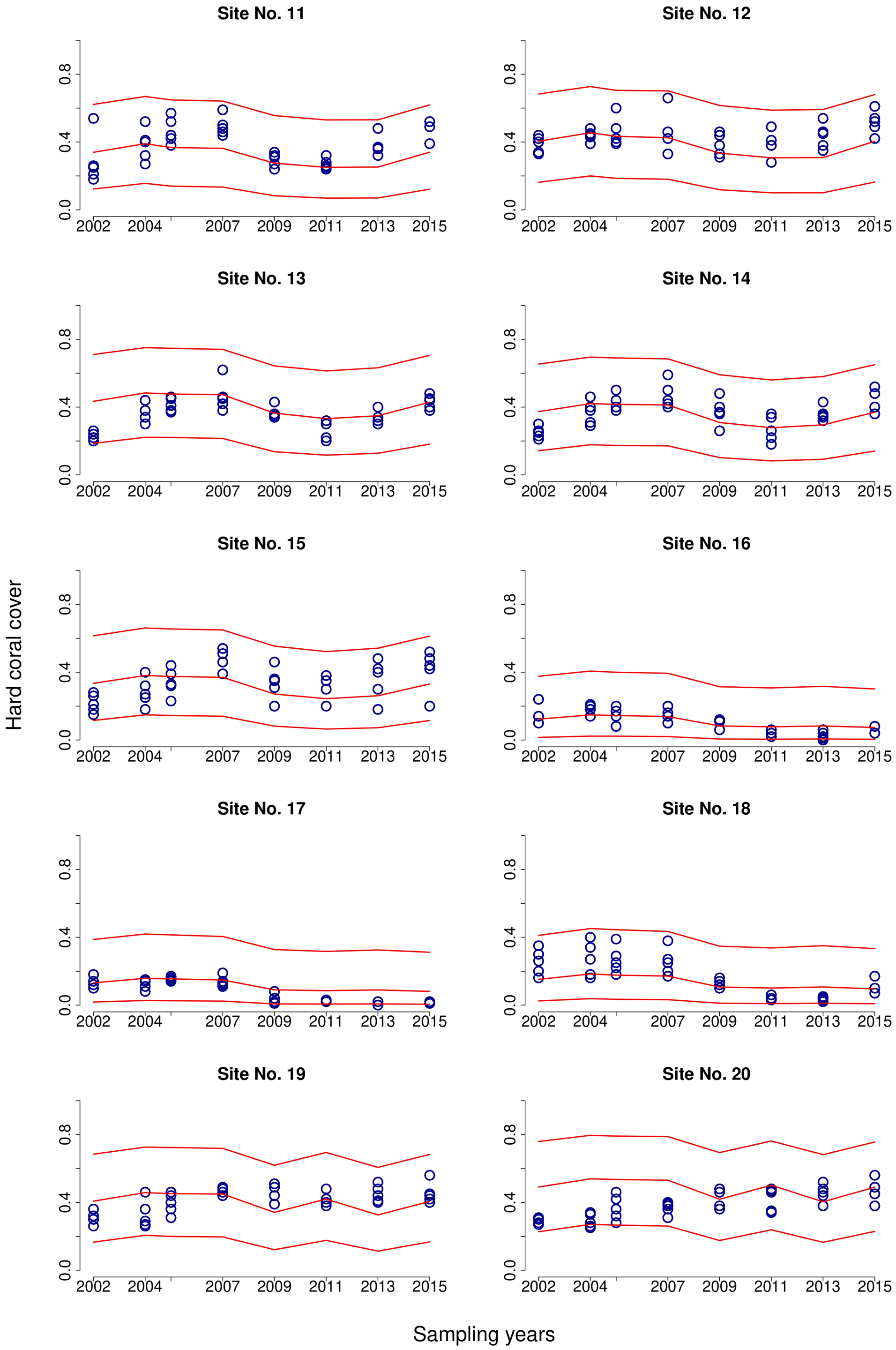

Fig. S3 Posterior predictive by sites for sites 11-20 in the Whitsunday region of the Great Barrier Reef. Red lines represent the median and the lower and upper bounds of 95% credible intervals for the prediction

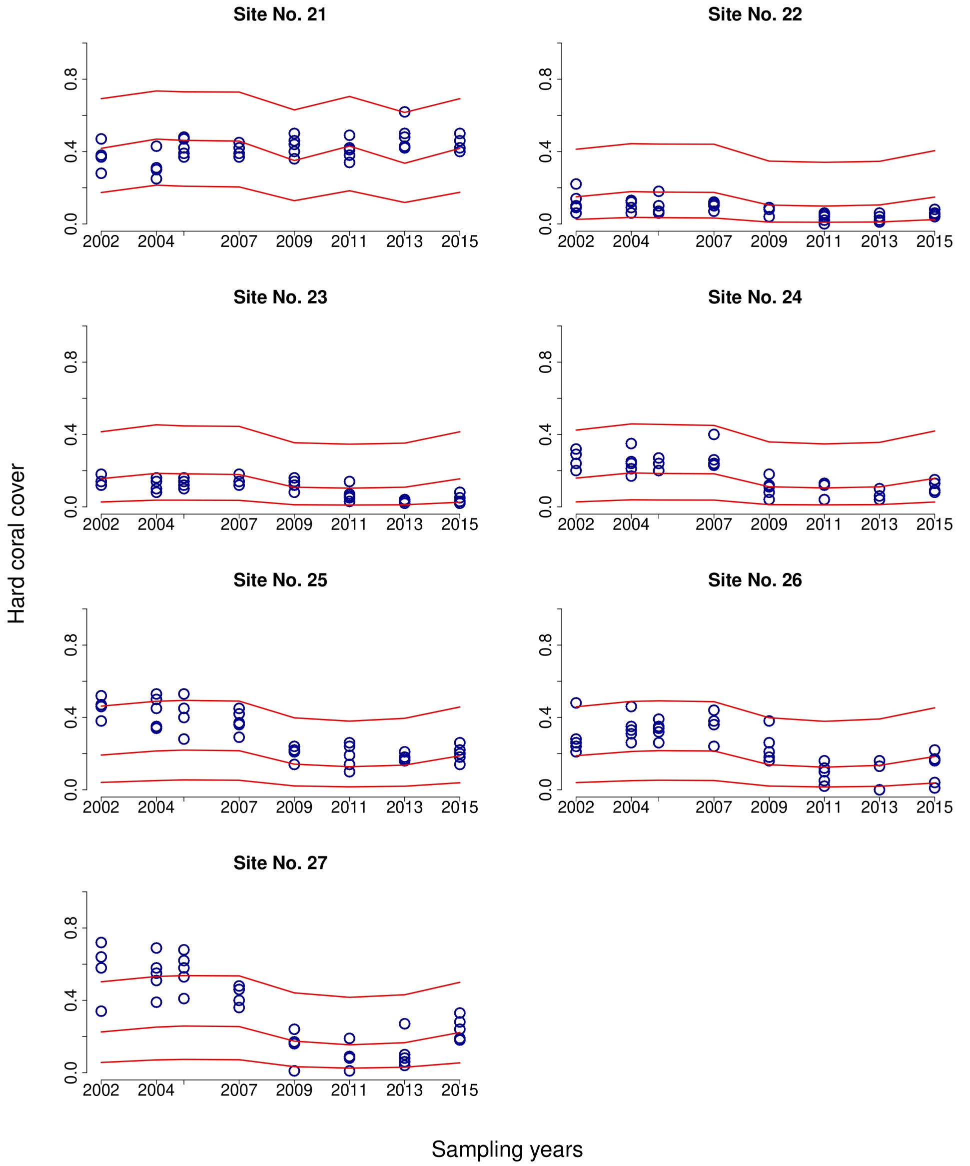

Fig. S4 Posterior predictive by sites for sites 21-27 in the Whitsunday region of the Great Barrier Reef. Red lines represent the median and the lower and upper bounds of 95% credible intervals for the prediction

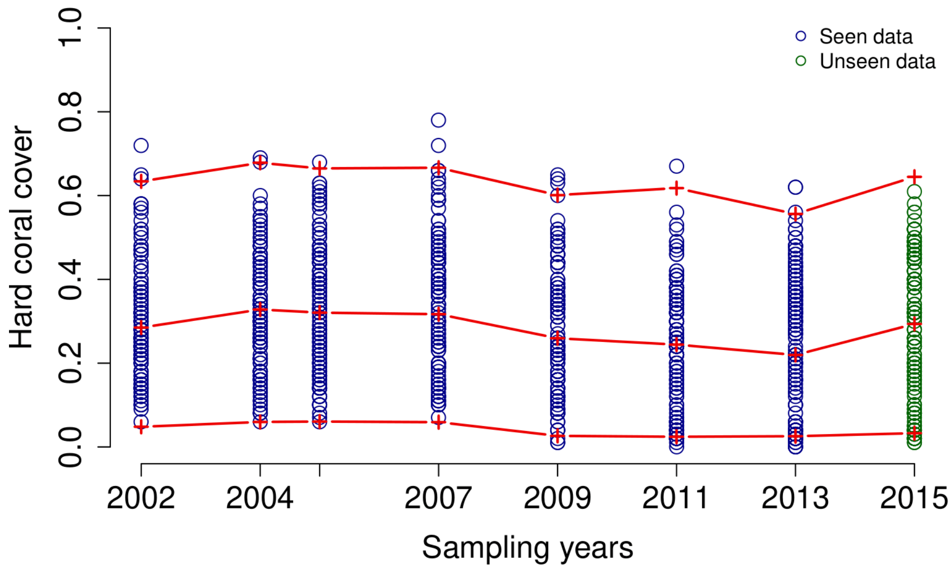

Fig. S5 Posterior predictive in the presence of "unseen" data. The blue and green circles represent "seen" and "unseen" data, respectively. The red lines represent the median and the lower and upper bounds for the 95% credible intervals for the prediction

References

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

Roberts DR, Bahn V, Ciuti S, et al (2017) Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography 40(8):913–929. https://doi.org/https://doi.org/10.1111/ecog.02881

Saltelli A, Tarantola S, Campolongo F, et al (2000) Sensitivity anaysis as an ingredient of modeling. Statistical Science 15(4):377–395. https://doi.org/https://doi.org/10.1214/ss/1009213004