MARINE MONITORING PROGRAM

Assessment of reproductive effort as an indicator of seagrass health for the Marine Monitoring Program

CSIRO
DATA 61
Australian Government
Great Barrier Reef
Marine Park Authority

Published by the Great Barrier Reef Marine Park Authority

ISBN 9780648357179

A catalogue record for this publication is available from the National Library of Australia

This publication should be cited as:

Lawrence, E., Gladish, D. 2019, *Assessment of reproductive effort as an indicator of seagrass health for the Marine Monitoring Program*. Great Barrier Reef Marine Park Authority, Townsville. 84 pp.

Front cover photo: Close up view of *Thalassia* hemp seagrass © Kor-jent van Dijk.

Comments and questions regarding this document are welcome and should be addressed to:

Great Barrier Reef Marine Park Authority
280 Flinders ST Townsville | PO Box 1379 Townsville QLD 4810
Phone: (07) 4750 0700
Fax: 07 4772 6093
Email: info@gbrmpa.gov.au
www.gbrmpa.gov.au

This project is supported by the Great Barrier Reef Marine Park Authority through its Marine Monitoring Program with funding from the Australian Government Reef Program, and the Australian and Queensland governments Reef 2050 Integrated Monitoring and Reporting Program.

Contents

Tables

Figures

Acknowledgments

We thank Katherine Martin and Carol Honchin from the Great Barrier Reef Marine Park Authority, for guidance and feedback during development and delivery of the project. We thank the Marine Monitoring Program providers for their advice and provision of data throughout this project. In particular, we thank Catherine Collier, Michelle Waycott and Len McKenzie for helping us understand the data and their valuable contributions on how to maximise the ecological relevance of the reproductive effort models. We would also like to acknowledge the members of the Reef 2050 Integrated Monitoring and Reporting Program Seagrass Working Group for the valuable discussions which led to improved project relevance and outputs.

Executive summary

We used models to determine whether seagrass data collected in the Marine Monitoring Program could be used to predict subsequent seagrass coverage. Parametric modelling (a traditional statistical modelling approach) along with a random forest modelling (a machine learning approach) was employed.

Three indicators were included as variables in the models: reproductive effort, species composition and prior levels of seagrass coverage. However, our primary objective was to assess how effective the reproductive effort was as a predictive indicator of seagrass condition and health. Reproductive effort includes male and female flowers, fruits, spathes, seeds and nodes.

While the reproductive effort data showed some promise in improving the predictions of seagrass coverage, the models often found weak associations due to the lack of statistical power in the data. Flowers were significant in all of the parametric models tested. Fruits and spathes were significant in some of the models indicating that they may have some value in predicting seagrass coverage. Overall, the high level of uncertainty associated with the reproductive effort index hinders the value of this indicator for assessing ecosystem health.

The abundance (percent cover) of seagrass in the previous year was the best predictor of seagrass cover at a site. The number of species was sometimes a good predictor of seagrass, although for some models once the indicators for the individual species were added, the number of species was no longer significant. These results support the assertion that species diversity and productivity are good indicators of resilience. The number of flowers featured as a key predictor in all the models and the number of fruit/spathes was important in the absence of lagged percent cover.

The presence of *Zostera* and *Cymodocea serrulata* were also key predictors of seagrass coverage, highlighting the importance of considering community composition.

While the seeds data was very important in the random forests, most likely the lack of power in this data prevented the seeds variable from being significant in the parametric models.

Continued collection of seeds data, and increasing the power in this dataset, would improve the understanding of how seed banks contribute to recovery in the Reef. The seed data should not be added to the reproductive effort metric at this stage as it would increase variability of an already low-powered metric.

Introduction

Reproductive effort is considered to be an indicator of the ability of a seagrass system to recover from loss (Collier and Waycott 2009). It is particularly important in the event of large-scale devastation such as caused by a storm or cyclone. McKenzie et al. 2016 attribute the recovery of seagrass meadows to reproductive output, seed banks and seagrass species composition.

The Marine Monitoring Program (MMP) currently reports reproductive effort based on a metric that incorporates all available information on the production of flowers and fruits per unit area (McKenzie et al. 2016). Given the high diversity of seagrass species and their variability in production of reproductive structures, this metric was seen as the best presentation of available information at the time of the program's inception. However, Kuhnert et al. 2015 reported that it is clear that of the three existing MMP indicators, reproductive effort has the least power and requires further investigation to determine how the metric could be improved for reporting.

The MMP collects information on seed banks for foundational seagrass species. The production of seeds is also considered to be a good indication of the ability of a meadow to recover from large scale impacts or even more localised losses (Collier and Waycott 2009). To date though, this information has not been included in the reproductive effort indicator due to perceived problems with the timing and representativeness of the data.

The roots and horizontal stems (rhizomes) of seagrass will often be beneath the sand/mud. Rhizomes are formed in segments and the leaves or vertical stems arising from the joins referred to as nodes. The MMP collect information on the number of nodes for each species, and this may be an indicator of the effort being put into growth/regeneration. This information is currently part of the supporting documentation for the MMP report and does not directly feed into an indicator.

Reproductive effort should not be analysed without giving consideration to community composition as the expected reproductive effort depends highly on the type of community and location. As highlighted by Udy et al. 2019, following loss, there is a large variation in recovery outcomes recorded in the Reef. This is likely a reflection of the degree to which meadows were initially impacted as well as differences in the availability of seagrass, seed banks or recruits to aid recovery.

The value of the data in predicting seagrass recovery, and the adequacy of the current MMP reproductive effort indicator, are tested and outcomes reported.

Data

The MMP collect data across a variety of habitat types in the Reef (coastal, estuarine and reef) using the methods described in McKenzie et al. 2016. The number of sites monitored has varied since inception but, for example, in the 2014-15 monitoring period 45 sites were monitored (typically two to three per location). The program also included additional data from the Seagrass-Watch program to improve the spatial resolution. Monitoring is undertaken in the late dry period (September/October) and late monsoon (March/April) of each year. At each site, observers record the percent seagrass cover within 33 quadrats (50 cm x 50 cm, placed every 5m along three 50m transects). The sampling strategy for subtidal sites is slightly modified to cater for the logistics of SCUBA diving in poor visibility.

Seagrass reproductive data is collected in the field during the MMP surveys. McKenzie et al. 2016 describe the sampling procedures in detail. In the field, 15 haphazardly placed cores (100mm diameter x 100mm depth) of seagrass are collected from an area adjacent to each monitoring site and the reproductive structures (spathes, fruits, female and male flowers) identified to species level and counted in the laboratory. Seed banks and abundance of germinated seeds are sampled by sieving 30 cores (50mm diameter, 100mm depth) of sediment collected across each site and counting the seeds retained in each. For *Zostera muelleri*, 18 cores are collected and returned to the laboratory where they were washed through a sieve and seeds identified using a hand lens/microscope.

Dominant species

The sites in the MMP have different species composition and pressures affecting them. Evidence of reproduction at a site will vary through time and depending on the timing of sampling, key reproductive effort indicators may vary from being in their peak to totally being absent (although having been present earlier). The peak season for reproduction for all species is considered to be between September and December.

We first determined the dominant species at each site and explored the number of nodes and reproductive effort data (flowers, fruit and spathes) for the dominant species at each site. We analysed each site separately, choosing not to aggregate to location. We plotted the percent cover and seeds per m^2 for all species at a site. We chose to plot combined total percent cover as the dominant species may change through time and a drop in cover of the normally dominant species may not necessarily represent a drop in overall cover. We plotted seed counts for all species as the data are sparse. As the date and number of samples can vary by site and year, we have grouped the data into ‘peak’ (Sept-Dec) and ‘non-peak’ (Jan-Aug) for each year and so each year where at least two samples have been collected will have two time points depicted (even if there were more). To give an indication of the trend in each of the variables at each site, we fitted a local polynomial regression smooth (loess). Note: the smooths are very simplistic and do not take seasonality into account, but still provide a general indication of the trend in the data.

The majority of sites have been dominated in the past by either *Halodule* or *Zostera* (Table 1). While we plotted the reproductive effort and seagrass cover data for all sites, we have chosen to show a subset of examples below so as not to overwhelm the reader, with the remaining sites in Appendix A. The examples cover a range of scenarios in the MMP. As reproductive effort is generally considered most critical during a major decline i.e. greater reproductive effort should result in a greater chance of recovery, we have mostly shown sites where there has been a major decline at some point. Some of the sites demonstrate a lot of evidence of reproduction after a decline (Site BB1) while others much less (Site GI1). The different responses are largely due to the environmental conditions, additional pressures and the dominant species at each site.

Table 1: Number of sites where each species is dominant (highest percent cover over time)

Species	**Number of sites**
Cymodocea rotundata (CR)	4
Cymodocea serrulata (CS)	2
Halophila ovalis (HO)	7
Halodule universis (HU)	24
Thalassia hemprichii (TH)	4
Zostera muelleri (ZM)	18

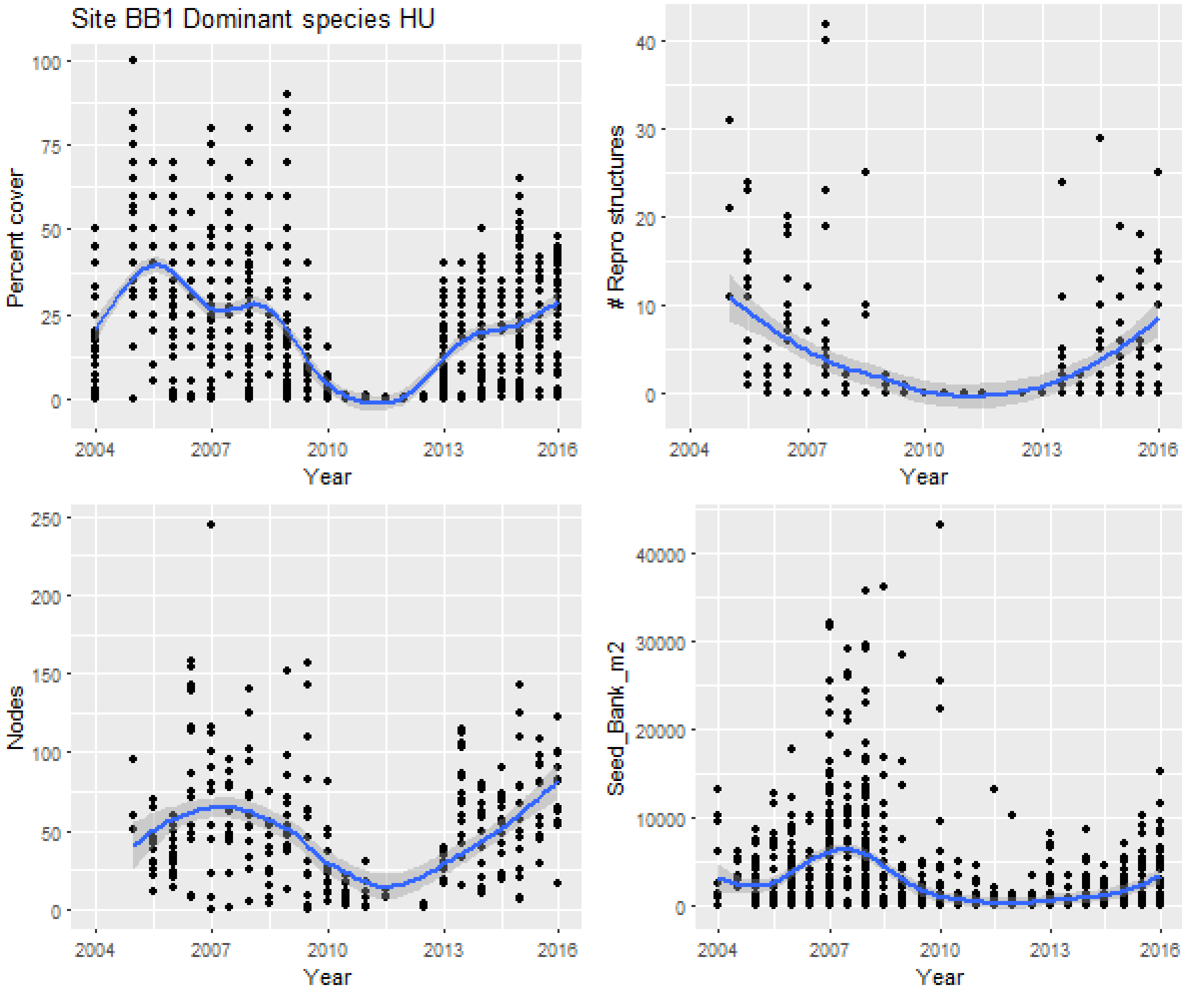

Figure 1:Summary of seagrass coverage and reproductive effort data at Bushland Beach (dominant species is Halodule)

Bushland Beach Site 1 experienced a big decline to almost zero cover around 2010 (Figure 1) but made a strong recovery from 2013 onwards. There is little evidence of reproductive structures during the years of major loss (as there is no seagrass to fruit/flower), however the structures come back quickly once the seagrass does. There is also evidence of nodes and seeds in the years of major loss, which likely would have contributed to the recovery of the species.

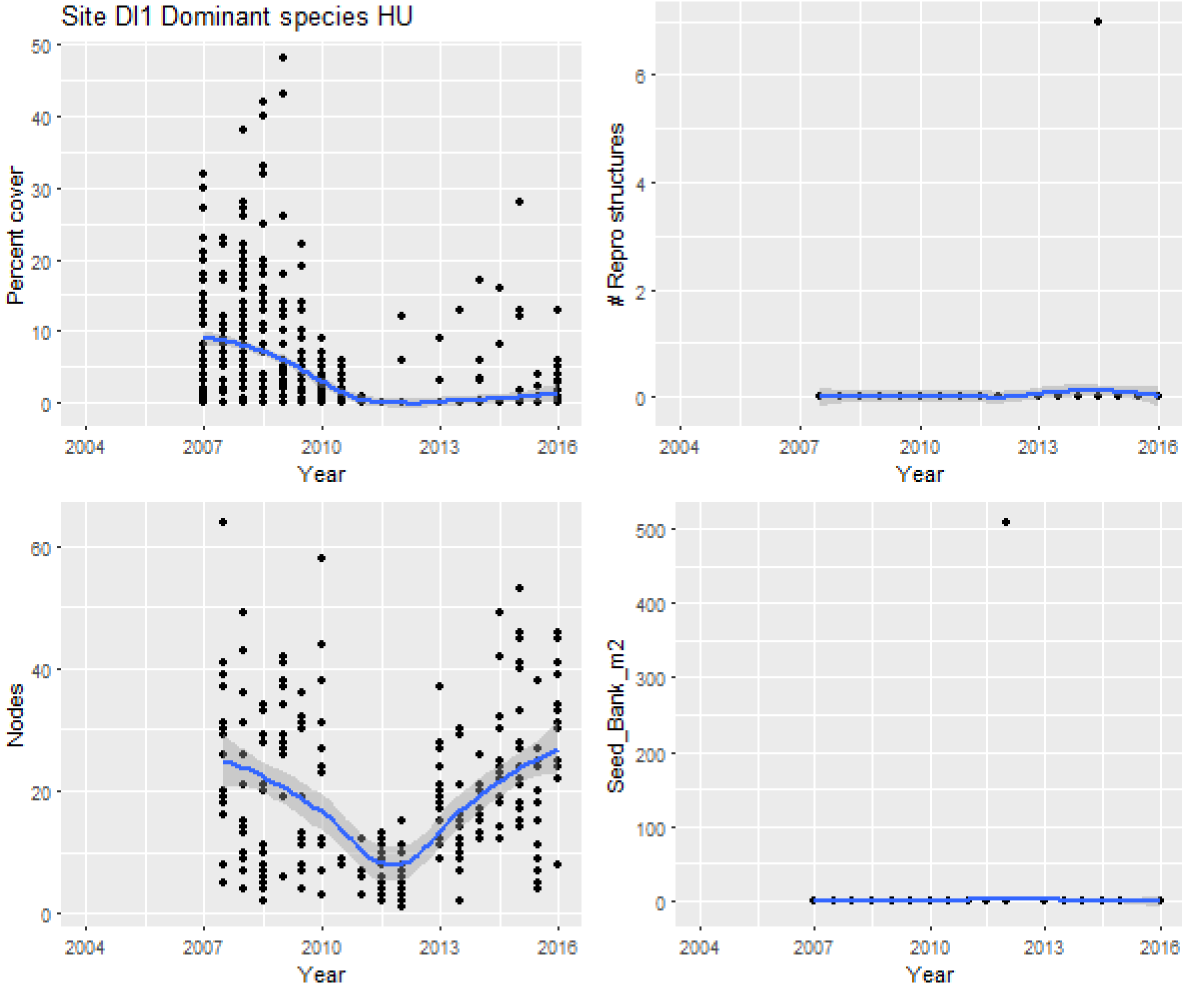

Figure 2: Summary of seagrass coverage and reproductive effort data at Dunk Island (dominant species is Halodule)

Dunk Island Site 1 exhibits a major loss in 2011 (Figure 2), with only a slight recovery afterwards. While almost no reproductive structures or seeds have been recorded at the site, there is an increasing number of nodes recorded over the last few years.

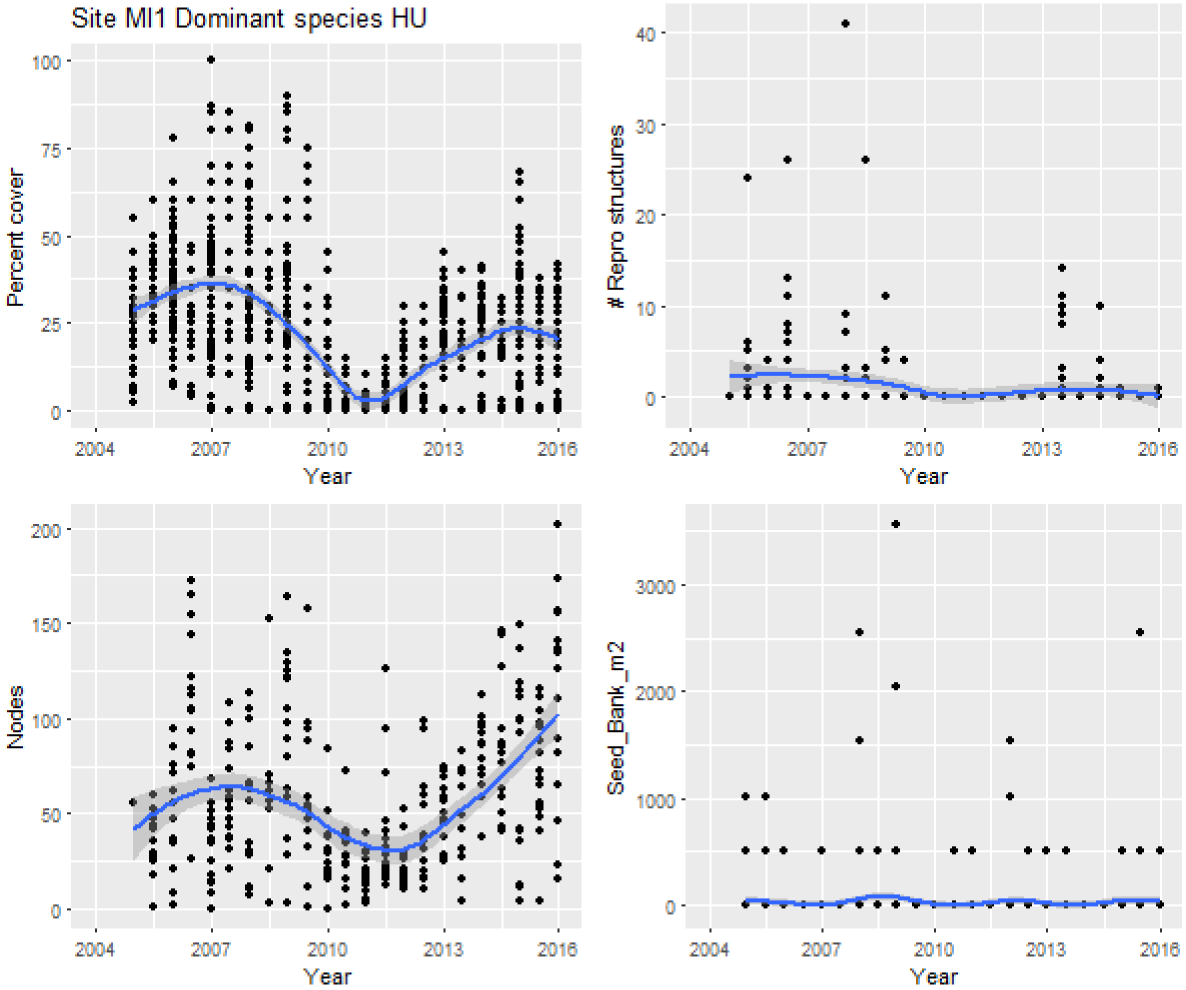

Figure 3: Summary of seagrass coverage and reproductive effort data at Picnic Bay Site 1 (dominant species is Halodule)

In 2011 Picnic Bay Site 1 also experienced a decline, although less marked than the previous two sites (Figure 3). Following the decline, the species recovers quickly and the increasing trend in nodes provides a good indication that the seagrass is working to recover. There is evidence of both reproductive structures and seed banks but no strong trends.

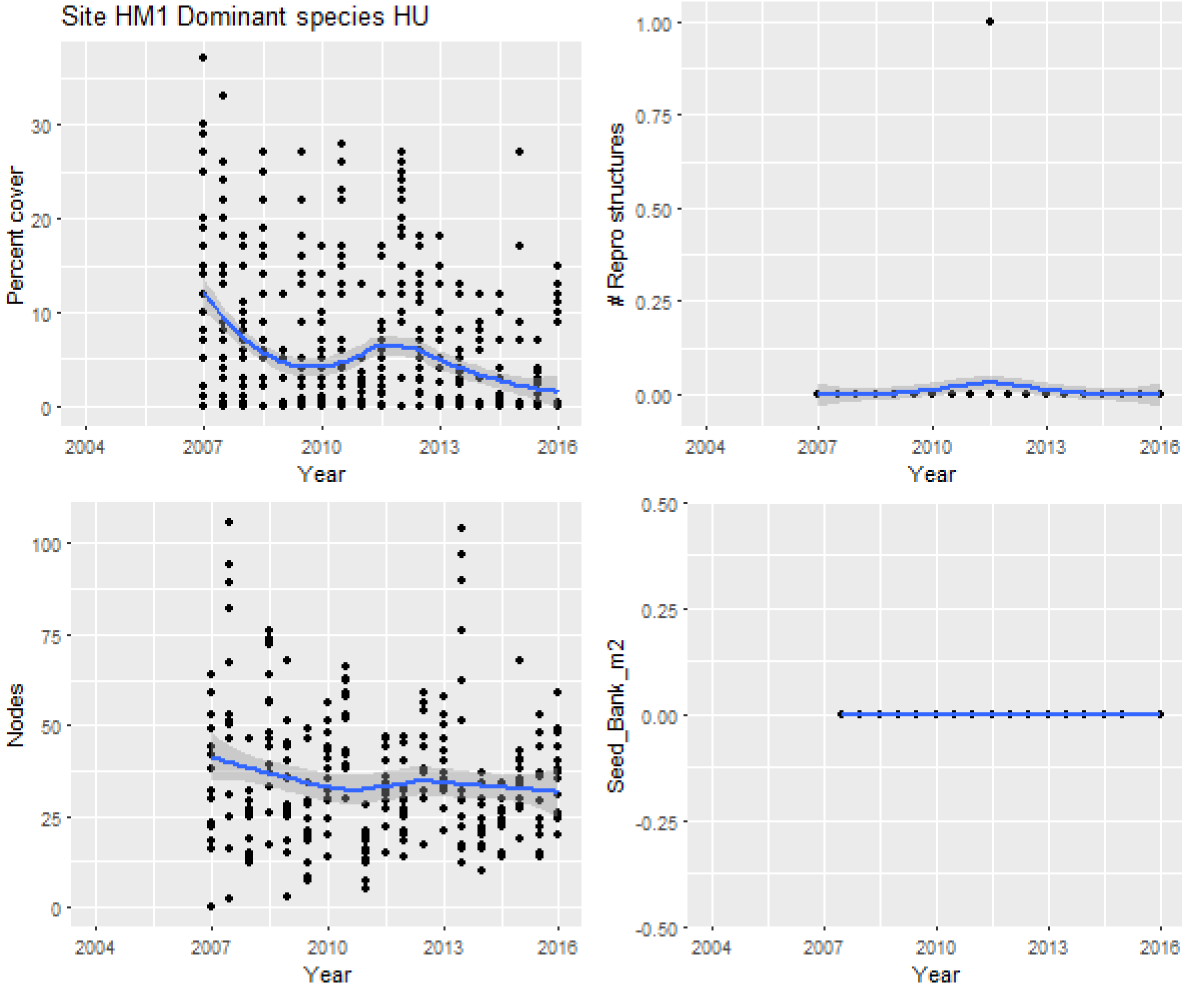

Figure 4: Summary of seagrass coverage and reproductive effort data at Hamilton Island Site 1 (dominant species is Halophila ovalis)

Hamilton Island Site 1 is an example of persistently fairly low seagrass coverage (Figure 4). There’s not much evidence to suggest that effort is being put in to “recover” or increase the standing biomass. This is a good example of how the expected abundance of seagrass, and effort put into recovery, differs by site.

Statistical Modelling Methods

One way to assess which environmental factors will act as key indicators in determining changes in percent cover is through statistical modelling. In this section, we outline the method for modelling seagrass as a function of environmental variables. Critically, interest is in which key variables significantly contribute to explaining percent cover of seagrass. We focus on developing two modelling frameworks: (i) a parametric Generalized Linear Model (GLM); and (ii) a non-parametric statistical model. Due to the fact that seagrass percent cover is bounded between 0 and 100%, with many of the observed values at or near zero percent cover, we aggregated to the site level first before fitting a Beta-distribution based generalized regression model and the non-parametric based Random Forest model.

Data Aggregation

Seagrass percent cover was observed at the quadrat level initially, and needed to be aligned with available seed and reproductive effort data that were collected using cores (not directly aligning with quadrats). We would like to determine key indicators at the site level over time. We focused on the peak/non-peak yearly time scale (as described earlier) at the site level. That is, we aggregated by taking the mean seagrass percent cover of all quadrats at a site for a given period (peak or non-peak season), per year.

For modelling purposes, we considered several seed bank and reproductive effort variables as well as site and time variables. Table 2 shows the variables of interest, namely seed data, factor variables dictating which site and which season is observed, and a variable that shows the percent cover of seagrass for the previous year for a given site and season. Importantly, the variables linked to reproduction (seeds, nodes, flowers, and spathes) are specified for the peak season only. That is, the observed reproductive effort variable values will be the same regardless of season for a given year and site. We also included indicator variables for each species to indicate presence/absence of each particular species as a means of considering community composition in our analysis.

Zero Inflated Beta Regression Modelling

We utilised a Generalized Linear Model (GLM) (Nelder and Wedderburn 1972) as a parametric based statistical framework to link potential indicators to seagrass percent cover. To accommodate the high percentage of zeros and the bounded nature of percent cover (which we scale to be between 0 and 1 inclusive), we used a zero-inflated beta distribution in the GLM framework (Ospina and Ferrari, 2010). The fine details can be found in Ospina and Ferrari (2010). We present a few technical details for reference here. We modeled percent cover of seagrass y as

$$y \sim Beta(\mu, \sigma, \nu) \tag{1}$$

where μ, σ, and ν are parameters associated with the mean, variation, and zero-inflation probability, respectively. These parameters are then specified as functions of the variables of interest. Specifically:

$$g_1(\mu) = \boldsymbol{X_1 \beta_1} \tag{2}$$

$$g_2(\sigma) = \boldsymbol{X_2 \beta_2} \tag{3}$$

$$g_3(\nu) = \boldsymbol{X_3 \beta_3} \tag{4}$$

where $\boldsymbol{X_1}$, $\boldsymbol{X_2}$, and $\boldsymbol{X_3}$ are matrices that contain explanatory variables for μ, σ, and ν, respectively, and $\boldsymbol{\beta_1}$, $\boldsymbol{\beta_2}$, and $\boldsymbol{\beta_3}$ are the respective vectors of covariate coefficients. Importantly, this formulation is flexible as it allows different potential indicators to influence the mean, variance, or the probability of non-zero percent cover. Critically, μ and ν relate to the response variables percent seagrass through a logit function, while σ relates through a log function.

Critically, key indicators will arise from associated covariates that are significant in the model. For simplicity, we specify the model g_1 of the mean parameter μ as a function of the environmental variables specified in Table 2, allowing the remaining functions g_2 and g_3 to be only modelled with an intercept term, though we note that future studies may want to investigate alternative forms for these functions. We fit this model in the R programming language (R Core Team, 2017) using the "gamlss" package (Rigby and Stasinopoulous 2005; Stasinopoulos et al. 2017).

Table 2: List of variables used for statistical modelling

Variable Name	**Description**
peak	Indicator variable dictating if in peak or non-peak season
num_species	Integer showing the number of species at the site for the season and year
Max_Nodes	Maximum number of nodes counted during one sampling point in the period
Max_Flowers	Maximum number of flowers counted during one sampling point in the period
Max_FruitsSpathes	Maximum number of fruits and spathes combined during one sampling point in the period
Seed_Bank_m2	Estimated sum of seeds per square meter
site	Factor variable indicating which Site
CR	Indicator for species *Cymodocea rotundata*
CS	Indicator for species *Cymodocea serrulata*
EA	Indicator for species *Enhalus acoroides*
HD	Indicator for species *Halophila decipiens*
HS	Indicator for species *Halophila spinulosa*
HO	Indicator for species *Halophila ovalis*
HU	Indicator for species *Halodule uninervis*
SI	Indicator for species *Syringodium isoetifolium*
TH	Indicator for species *Thalassia hemprichii*
ZM	Indicator for species *Zostera muelleri*
Lagged_Cover	A variable indicating the previous year mean percent cover

Model Formulation

Both the zero-inflated beta regression model and the random forest model link the explanatory variables in Table 2 to the percent cover of seagrass. In this section, we present the details of the formulations we use for our analysis. We consider two different specifications of percent cover of seagrass and four alternative models.

The four models can be written as:

$$\begin{aligned} \text{M1: } f(y_{ijk}) = \beta_0 &+ \text{peak}_k + \text{num\_species}_{ij-1k} + \text{Nodes}_{ij-1} + \text{MFlowers}_{ij-1} + \text{FFlowers}_{ij-1} \\ &+ \text{Spathes}_{ij-1} + \text{Seed\_Bank\_m2}_{ij-1k} + \text{site}_i + e_{ijk}, \end{aligned}$$

$$\begin{aligned} \text{M2: } f(y_{ijk}) = \beta_0 &+ \text{peak}_k + \text{num\_species}_{ij-1k} + \text{Nodes}_{ij-1} + \text{MFlowers}_{ij-1} + \text{FFlowers}_{ij-1} \\ &+ \text{Spathes}_{ij-1} + \text{Seed\_Bank\_m2}_{ij-1k} + \text{site}_i + \text{CR}_{ij-1k} + \text{CS}_{ij-1k} + \text{EA}_{ij-1k} \\ &+ \text{HD}_{ij-1k} + \text{HS}_{ij-1k} + \text{HO}_{ij-1k} + \text{HU}_{ij-1k} + \text{SI}_{ij-1k} + \text{TH}_{ij-1k} + \text{ZM}_{ij-1k} + e_{ijk}, \end{aligned}$$

$$\begin{aligned}\text{M3}: f(y_{ijk}) = {} & \beta_0 + \text{peak}_k + \text{num\_species}_{ij-1k} + \text{Nodes}_{ij-1} + \text{MFlowers}_{ij-1} + \text{FFlowers}_{ij-1} \\ & + \text{Spathes}_{ij-1} + \text{Seed\_Bank\_m2}_{ij-1k} + \text{site}_i + \text{lagged\_cover}_{ijk} + e_{ijk},\end{aligned}$$

$$\begin{aligned}\text{M4}: f(y_{ijk}) = {} & \beta_0 + \text{peak}_k + \text{num\_species}_{ij-1k} + \text{Nodes}_{ij-1} + \text{MFlowers}_{ij-1} + \text{FFlowers}_{ij-1} \\ & + \text{Spathes}_{ij-1} + \text{Seed\_Bank\_m2}_{ij-1k} + \text{site}_i + \text{CR}_{ij-1k} + \text{CS}_{ij-1k} + \text{EA}_{ij-1k} \\ & + \text{HD}_{ij-1k} + \text{HS}_{ij-1k} + \text{HO}_{ij-1k} + \text{HU}_{ij-1k} + \text{SI}_{ij-1k} + \text{TH}_{ij-1k} + \text{ZM}_{ij-1k} \\ & + \text{lagged\_cover}_{ijk} + e_{ijk},\end{aligned}$$

where i is the index referencing site, j is the index referencing ‘financial’ year (July-June), k is the index referencing season (either peak or non-peak season), y_{ijk} represents the chosen percent cover measurement at site i, year j, season k, β_0 represents the intercept term, and e_{ijk} represents the associated error. We use $f()$ as a generic term in M1-M4 to represent either the zero-inflated beta regression or random forest specification described above.

In summary terms:

- M1 is predicting seagrass percent cover based on current season, number of species present in the previous year, the reproductive/resilience variables from the peak of the previous year (nodes, flowers, fruits, spathes and seeds).
- M2 is the same but with indicator variables showing which species were present in the previous year.
- M3 is the same as M1 but with the addition of the seagrass cover in the previous year.
- M4 is the same as M3 with the addition of the species indicator variables.

We further considered two definitions of percent cover. The first was mean total cover based on all species, which we label as $\text{percent\_cover}_{ijk}$. The second was mean combined cover of *Halodule* and *Zostera*, which we label as hu_zm_cover_{ijk}. Both model types rely on having values for all variables i.e. no NA values. If any variable has an NA in the model, that observation is omitted in the analyses.

Statistical Modelling Results

We applied the four model formulations, fitting both a zero-inflated beta regression and a random forest statistical model to the seagrass data for the MMP. In doing so, we are looking for which variables significantly contribute to explaining percent cover of seagrass.

After fitting the zero-inflated beta regression and random forest for specifications M1-M4, for both the full seagrass percent cover and *Halodule/Zostera* percent cover, we tested how well the models fit the observed data. We did this by fitting a simple linear regression to the fitted value against the observed and obtained the respective R-squared values. Table 3 shows these R-squared values. All formulations fit the data reasonably well for both statistical models and both percent cover values, with R-squared values between 0.50 and 0.74. We note that the R squared values for the zero-inflated beta model is always at least slightly better than the random forest models, although the increase is sometimes very modest. The R squared values also increase moving from M1 to M4 indicating that the models are improved by considering both species composition and the amount of seagrass in the year prior.

Table 3: R squared values for a linear regression of the fitted values against the observed value for for models M1-M4 when fit with a zero-inflated beta regression or a random forest to percent cover of all seagrass species

Model	**Zero-inflated beta**	**Random Forest**
M1	0.653	0.633
M2	0674	0.652
M3	0.709	0.673
M4	0.738	0.684

Table 4: R squared values for a linear regression of the fitted values against the observed value for for models M1-M4 when fit with a zero-inflated beta regression or a random forest to percent cover of Halodule and Zostera

Model	**Zero-inflated beta**	**Random Forest**
M1	0.550	0.508
M2	0.570	0.515
M3	0.628	0.537
M4	0.668	0.567

As the M4 models gave the best R squared values, we predicted from both M4 models and compared the predictions to the observed values at each site (Figure 5). A subset of sites are shown and the rest can be found in Appendix B. Where the observed values are missing it is because those values have not been used in the model, usually due to missing reproductive effort variables. For most sites the models do quite well, even predicting recovery following decline for sites such as Bushland Beach (although the beta regression model provides much better predictions at this site). At a small number of sites, such as the two in Gladstone Harbour, the models are much further from the observed values though.

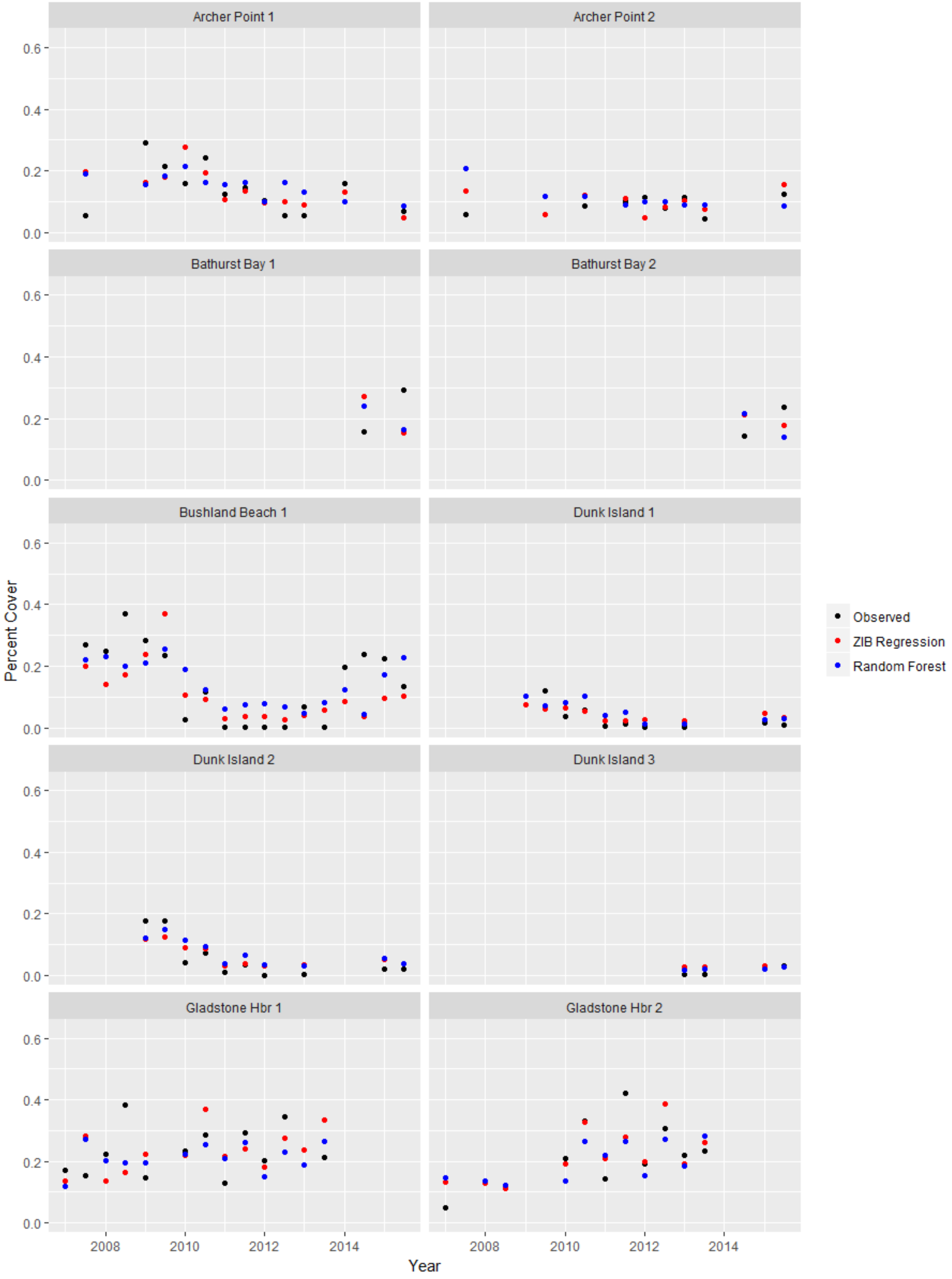

Figure 5: Observed (black) and predicted percent cover of seagrass for the zero inflated beta regression (red) and random forest (blue) for individual sites using model M4

We also considered the predictive capabilities of both the zero-inflated beta regression and the random forest. Our approach was to test how well each model does in predicting 2012. To do this, we fitted both statistical models using data at all available sites from 2006 through 2011, and used the resulting model fit to predict seagrass cover for both peak and non-peak seasons in 2012 at all sites that have available data. We then computed the mean squared error (MSE) of the predicted seagrass cover. Table 5 and Table 6 shows the MSE values. The zero-inflated beta models consistently outperform the equivalent random forests models. We chose 2012 as a particularly difficult year to predict given many sites were recovering following a decline but by choosing this year we effectively omitted 5 years of data from the "training" of the model so we expect future predictions to be an improvement. The zero inflated beta model may also be able to be improved slightly with further work to estimate the other two parameters that we have held fixed at this stage.

Table 5: Predicted mean squared error (MSE) of the zero-inflated beta regression and random forest models for formulations M1-M4 for total percent cover

Model	**Zero-inflated beta**	**Random Forest**
M1	0.00759	0.00944
M2	0.00822	0.0097
M3	0.00459	0.00637
M4	0.00518	0.0068

Table 6: Predicted mean squared error (MSE) of the zero-inflated beta regression and random forest models for formulations M1-M4 for Halodule and Zostera percent cover

Model	**Zero-inflated beta**	**Random Forest**
M1	0.0057	0.00648
M2	0.00549	0.00668
M3	0.00322	0.0049
M4	0.00357	0.00466

Zero-Inflated Beta Regression Key Indicators

Given the results indicating a good model fit, we determined which variables are potentially key indicators of seagrass cover. We did this by examining the associated p-values of the resulting estimated coefficients of each variable, for each formulation M1-M4. Table 7 shows the estimated coefficients and p-values for each formulation. There are a lot of interesting relationships highlighted through the results of these models, so we have summarised our findings:

- The number of species in the year prior is significant in the models that don't have the individual species indicators or lagged cover (M1, M2 and M3).
- The number of nodes in the year prior is significant in M4. The coefficients are negative indicating that a higher number of nodes in the year prior results in less seagrass. This could be due to the seagrass "working hard" to regenerate in periods of loss.
- The number of flowers was significant in all of the models. In each case, the greater the number of flowers in the year prior, the greater the mean seagrass prediction.
- The number of combined fruits/spathes was significant for M1 and M2 but not once lagged cover was added.
- The seed bank variable was never significant.

- *Cymodocea serrulata and Zostera* were significant in both of the models where they were considered. Each had a positive coefficient indicating that the presence of these seagrass in a particular year is an indicator for higher mean seagrass the following year. These are all considered to be foundation species in the Reef.
- Lagged cover was always significant when included in the models.

We haven't tried to interpret the coefficients, other than their direction, as the model is complex and a logit link was applied meaning that the relationships are non-linear.

Table 7: Coefficients and p-values for the Zero-inflated beta regression for percent_cover response under formulations M1-M4

Variable	**M1**	**M2**	**M3**	**M4**
Peak yes	0.085 (0.22)	0.052 (0.452)	-0.002 (0.974)	-0.046 (0.48)
num_species	**0.306** **(0.000)**	**0.253** **(0.017)**	**0.265** **(0.000)**	0.147 (0.135)
Max_Nodes	-0.001 (0.05)	-0.001 (0.06)	-0.001 (0.088)	**-0.001** **(0.043)**
Max_Flowers	**0.0207** **(0.000)**	**0.020** **(0.000)**	**0.015** **(0.001)**	**0.014** **(0.001)**
Max_FruitsSpathes	**0.013** **(0.001**	**0.014** **(0.000)**	0.005 (0.253)	0.006 (0.101)
Seed_Bank_m2	0.000 (0.297)	0.000 (0.393)	0.000 (0.671)	0.000 (0.827)
CR	NA	0.141 (0.556)	NA	0.07 (0.975)
CS	NA	**0.375** **(0.026)**	NA	**0.471** **(0.003)**
EA	NA	-0.279 (0.587)	NA	-0.786 (0.118)
HD	NA	-0.162 (0.768)	NA	0.105 (0.838)
HS	NA	-0.282 (0.603)	NA	-0.954 (0.065)
HO	NA	-0.000 (0.995)	NA	0.156 (0.266)
HU	NA	-0.119 (0.473)	NA	0.003 (0.986)
SI	NA	-0.115 (0.779)	NA	-0.192 (0.637)
TH	NA	-0.153 (0.365)	NA	-0.132 (0.396)
ZM	NA	**0.491** **(0.003)**	NA	**0.491** **(0.001)**
lagged_cover	NA	NA	**3.271** **(0.000)**	**3.570** **(0.000)**

The p-value is in the brackets and bold indicates p-value less than 0.05.

We also considered a second response variable, the percent cover based on the combined *Zostera* and *Halodule* cover. This analysis was performed separately to determine whether the number of seeds became an important predictor when only considering those species where seeds can most accurately be measured. Note that although we only included the combined cover of the two species, we still left the indicators regarding the presence of other species in the model.

Summarising our findings:

- The number of species in the year prior is significant in the models that don't have the individual species indicators (M1 and M3).
- The number of nodes in the year prior is significant in M1, M3 and M4. The coefficients are negative indicating that a higher number of nodes in the year prior results in less seagrass. This could be due to the seagrass "working hard" to regenerate in periods of loss.
- The number of flowers was significant in all of the models. In each case, the greater the number of flowers in the year prior, the greater the mean seagrass prediction.
- The number of combined fruits/spathes was significant for M1 and M2 but not once lagged cover was added.
- The seed bank variable was never significant.
- The presence of *Zostera* in the year prior is a good indicator for *Zostera* cover in a given year ($p < 0.05$). *Thalassia* and *Halophila Spinulosa* were both significant in M4 but not M3.
- Lagged cover was always significant when included in the models.
- While not given in Table 8, the model coefficients were similar in magnitude to those for the percent cover model.

There is little difference in the results here compared to the models for total percent cover and we note that the seed bank variable is still not significant when we restrict to just Zostera/Halodule cover.

Table 8: P-values for the Zero-inflated beta regression for hu_zm_cover response under formulations M1-M4

Variable	**M1**	**M2**	**M3**	**M4**
Peak yes	0.349	0.660	0.543(-)	0.099(-)
num_species	**0.000**	0.099	**0.001**	0.488
Max_Nodes	**0.025(-)**	**0.021(-)**	0.052(-)	**0.020(-)**
Max_Flowers	**0.000**	**0.000**	**0.001**	**0.000**
Max_FruitsSpathes	**0.001**	**0.000**	0.483	0.221
Seed_Bank_m2	0.595	0.742	0.973	0.547(-)
CR	NA	0.425	NA	0.482
CS	NA	0.253	NA	0.053
EA	NA	0.915(-)	NA	0.195(-)
HD	NA	0.446(-)	NA	0.780(-)
HS	NA	0.652(-)	NA	**0.046(-)**
HO	NA	0.636	NA	0.153
HU	NA	0.750(-)	NA	0.410(-)
SI	NA	0.809(-)	NA	0.695(-)
TH	NA	0.366(-)	NA	**0.026**
ZM	NA	**0.001**	NA	**0.000**
lagged_cover	NA	#NA	**0.000**	**0.000**

The p-value is in the brackets and bold indicates p-value less than 0.05. Note the estimated model coefficients are not given but the (-) indicates a negative coefficient.

Random Forest Key Indicators

We next present the results for the random forest models. Since random forests are non-parametric in nature, we rank the importance of the variable rather than their indicating significance. Importance is based on what is known as reduction in MSE which indicates the reduction in error achieved through the introduction of each variable, with the variables depicted in decreasing value (importance) (see Breiman, 2001 for details). Figures 6-9 show the importance rankings for percent_cover for M1-M4 while Figures 10-13 show the importance rankings for the combined cover of *Halodule* and *Zostera*.

As with the zero-inflated beta regression, the number of species, nodes, number of flowers, number of fruits/spathes and lagged cover were consistently important predictors of percent cover. The two *Cymodocea* (CR and CS) species and *Zostera* also showed importance when species were added, supporting the zero-inflated beta regression results*. Zostera* was the most important species variable for hu_zm_cover, agreeing with the zero-inflated beta regression results.

Interestingly, seed_bank_m2 is important in the random forest model, disagreeing with the zero-inflated beta regression. This is likely to mean that seeds are important in predicting seagrass, and in particular seagrass recovery, but due to the sparsity of the data there is not enough power to display significance in the beta regression model. As such, while seed_bank_m2 may be too variable to be a good indicator of seagrass cover currently (due to low seed numbers), it may still be valuable in predicting seagrass recovery when combined with other relevant variables. This is a slight but critical distinction to note.

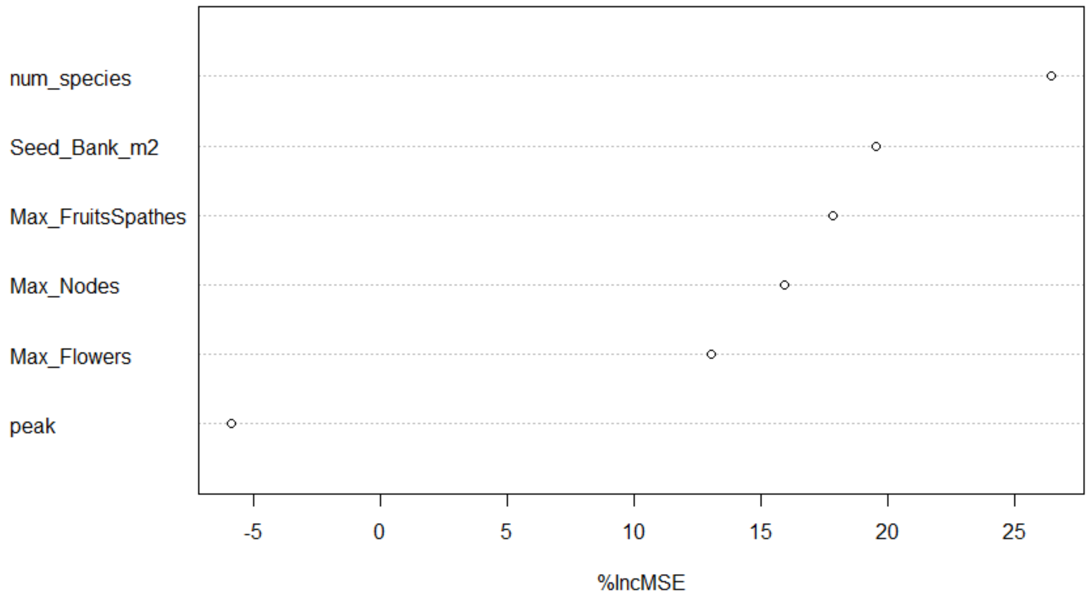

Figure 6: Variability importance ranking of the random forest model for M1 considering the entire percent_cover

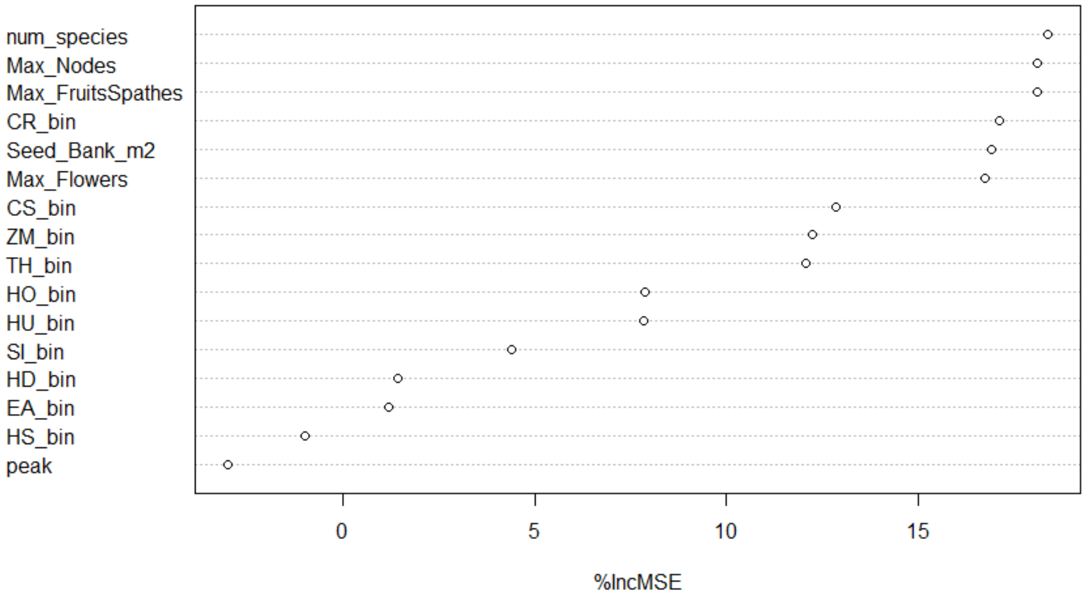

Figure 7: Variability importance ranking of the random forest model for M2 considering the entire percent_cover

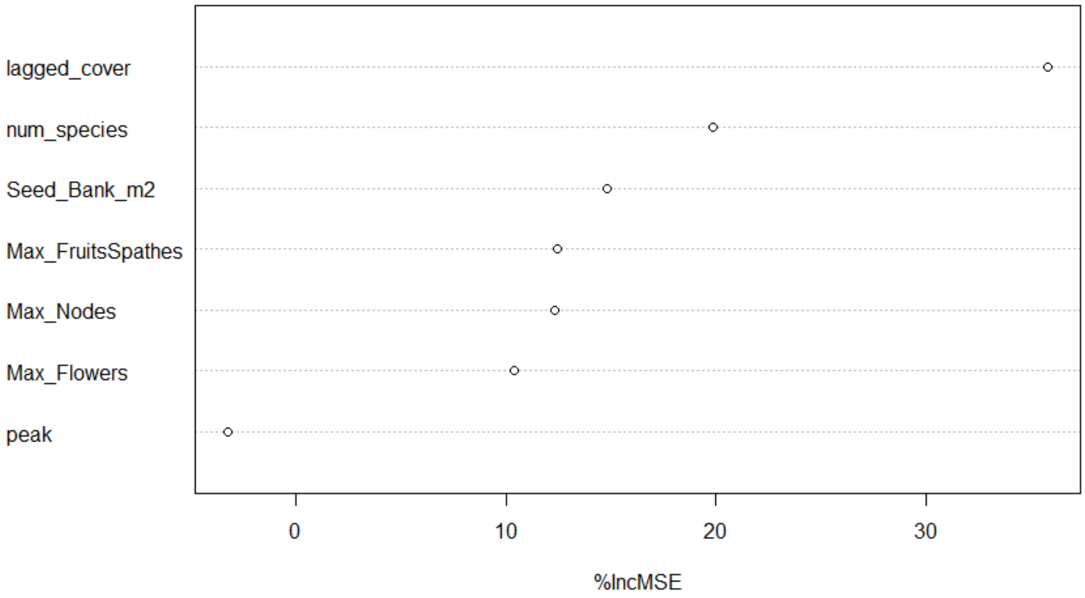

Figure 8: Variability importance ranking of the random forest model for M3 considering the entire percent_cover

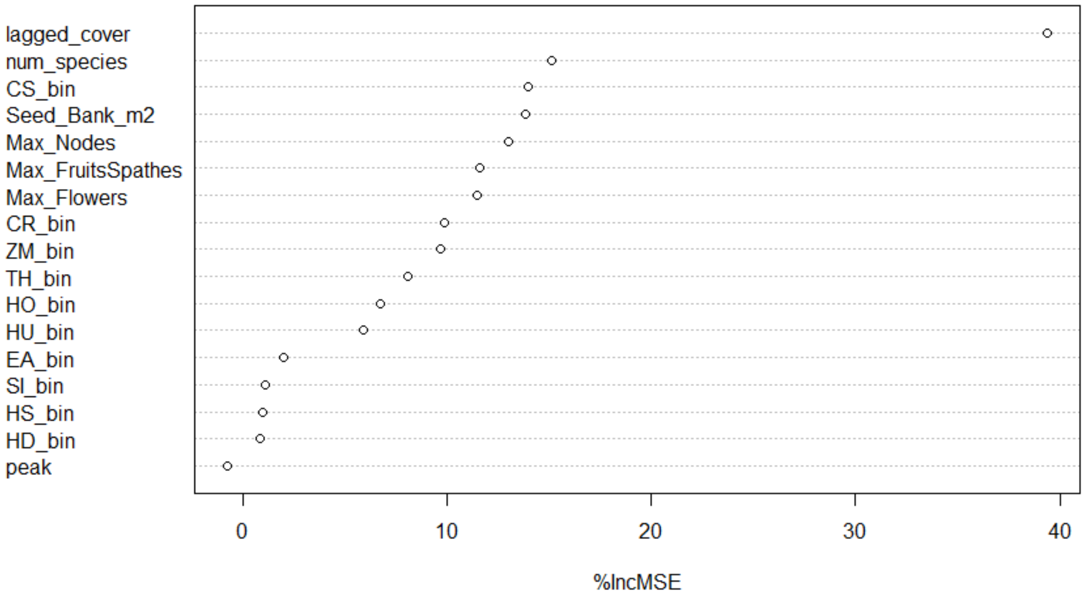

Figure 9: Variability importance ranking of the random forest model for M4 considering the entire percent_cover

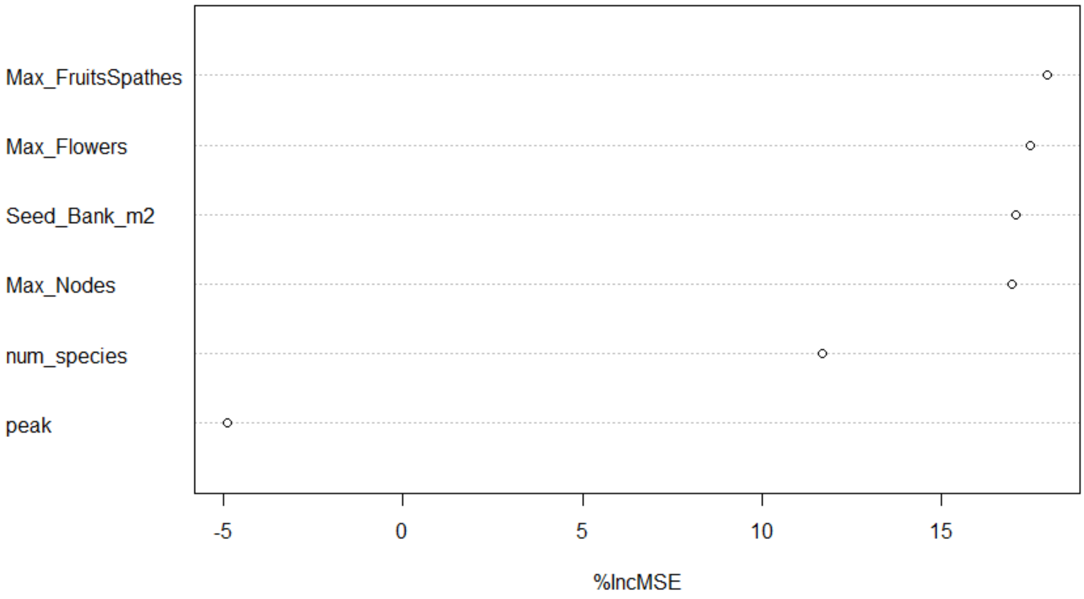

Figure 10: Variability importance ranking of the random forest model for M1 considering weighted average of HU and ZM percent cover

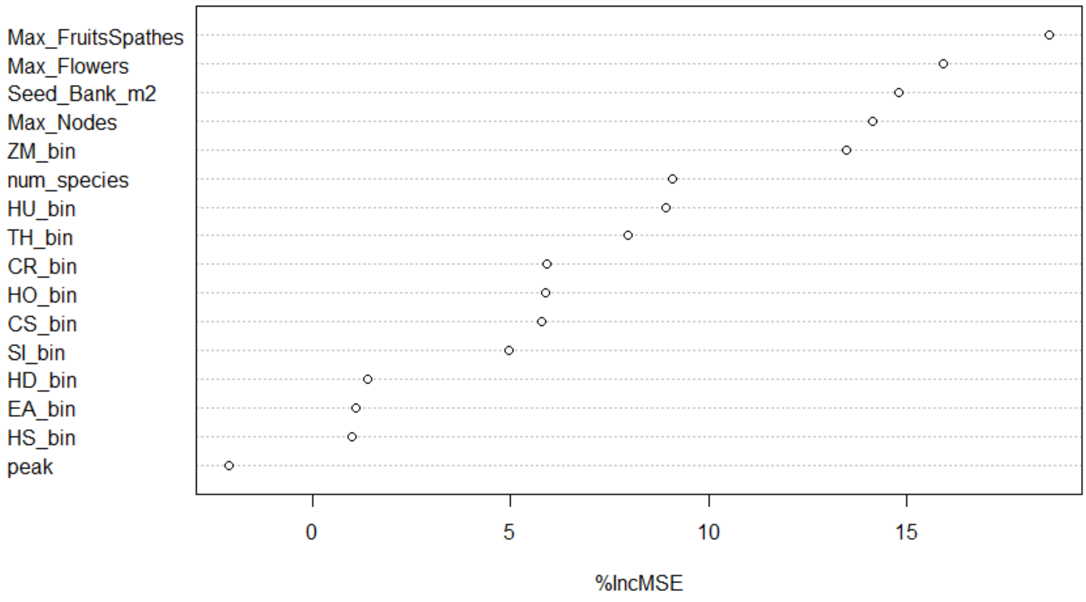

Figure 11: Variability importance ranking of the random forest model for M2 considering weighted average of HU and ZM percent cover

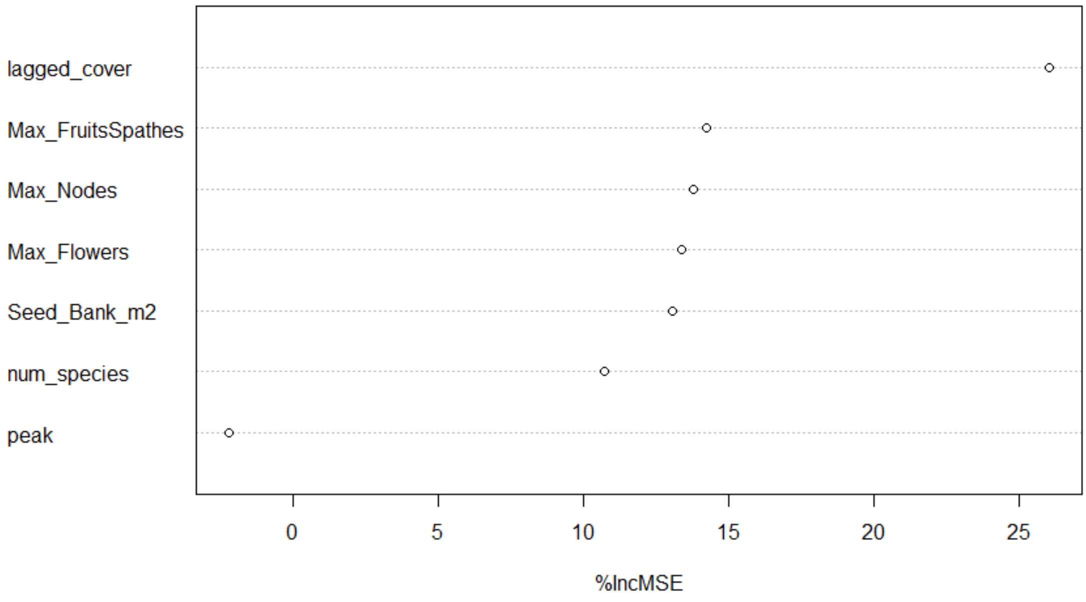

Figure 12: Variability importance ranking of the random forest model for M3 considering weighted average of HU and ZM percent cover

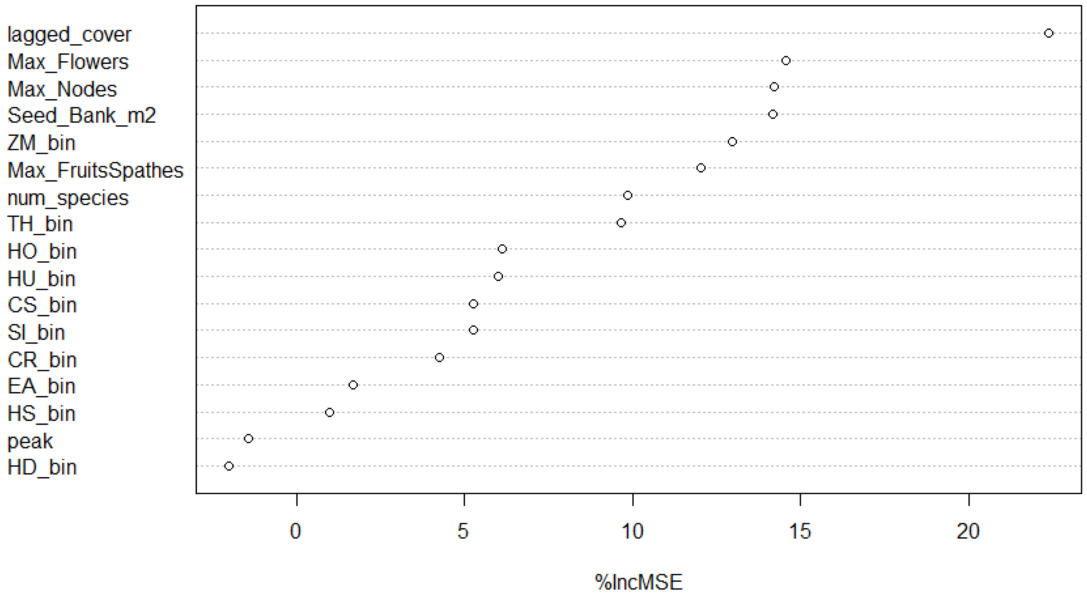

Figure 13: Variability importance ranking of the random forest model for M4 considering weighted average of HU and ZM percent cover. Importance ranking is based on node purity

Discussion and implications for the MMP and future monitoring

The sites in the MMP vary in their species composition, baseline seagrass coverage, levels of reproductive effort and the pressures affecting them. We tried to predict seagrass coverage based on reproductive effort, species composition and prior levels of coverage, for all sites in the MMP where the amount of data was adequate to do so. We used both a parametric modelling and a random forests approach, with both producing similar levels of accuracy and a similar set of important predictor variables. While random forests are an excellent tool for prediction, the beta regression models provided consistently better accuracy and have the advantage of providing better statistical inference and relationships between variables that are understandable.

The amount of seagrass in the previous year is unsurprisingly the best predictor of seagrass coverage at a site. The number of species was also a good predictor of seagrass, however once the indicators for the individual species were added, the number of species was no longer significant. These results support the assertion that species diversity and productivity are good indicators of resilience (Unsworth et al. 2015). The number of flowers featured as a key predictor in all of the models and the number of fruits/spathes was important in the absence of lagged percent cover.

In terms of community composition, those sites with *Zostera* and *Cymodocea serrulata* in the previous year were associated with higher amounts of seagrass coverage. These results confirm the need to consider community composition when considering the likely time to recover following a significant loss.

While the reproductive effort data showed some promise in improving the predictions of seagrass coverage, the models typically found weak associations due to the power in the data. Flowers, fruits and spathes (the variables currently forming the MMP reproductive effort indicator) were all significant in some of the models indicating that they have some value in predicting seagrass coverage and are therefore useful variables to continue to monitor and report on. While the seeds data came out as very important in the random forests, most likely the lack of power in this data prevented this variable from being significant in most of the parametric models. We would recommend to continue to collect the seeds data and increase the power in this dataset where possible to improve the understanding of how seed banks contribute to recovery in the Reef. We would not recommend adding it to the reproductive effort metric at this stage as it would add further variability to an already highly uncertain metric.

While we have demonstrated that the reproductive effort measures currently aggregated to form the MMP reproductive index are useful in predicting seagrass abundance, the power analysis conducted by Kuhnert et al. 2015 raised concerns over the amount of power to detect change using this metric. The model results support this assertion. Below we plot the reproductive effort indicator and associated confidence interval by site in each of the habitat categories to provide an indication of the uncertainty associated with this metric (Figure 14 - Figure 18). Due to the differences between some sites within a location, we have plotted each site separately but note in some cases combining them would be warranted and reduce the associated uncertainty.

The sum of the reproductive effort measures forms a count variable and due to the scarcity of the data at some sites and times, the data are zero-inflated meaning that there is a greater proportion of zero’s than would be expected for a typical count process. For this reason the standard errors are likely to be underestimated by reporting the standard normal (Gaussian) standard error currently reported by the MMP. To give a more accurate estimate of uncertainty we fitted a simple negative binomial GLM model to the data, to account for the excess zero's and high variability, and used the standard errors estimated by the model. The model requires more data to calculate the estimates than a typical standard error calculation and for that reason the confidence interval estimates are missing from some of the plots, although a mean is reported.

The figures also show the 'very poor', 'poor', 'moderate' and 'good' categories for the reproductive effort indicator. Here it is apparent that the high level of uncertainty associated with this index is hindering the adequacy of this metric. The confidence intervals at many of the sites cross three and sometimes four categories, for example at Site BB1 in both 2006 and 2016 the reproductive effort indicator is reported as being 'moderate' but the confidence intervals range from 'very poor' to 'good'. The difference between being 'very poor' reproductive effort and 'good' reproductive effort would most likely have a large impact on the ability of a meadow to recover in the event of a large loss, but the metric can't differentiate between these categories.

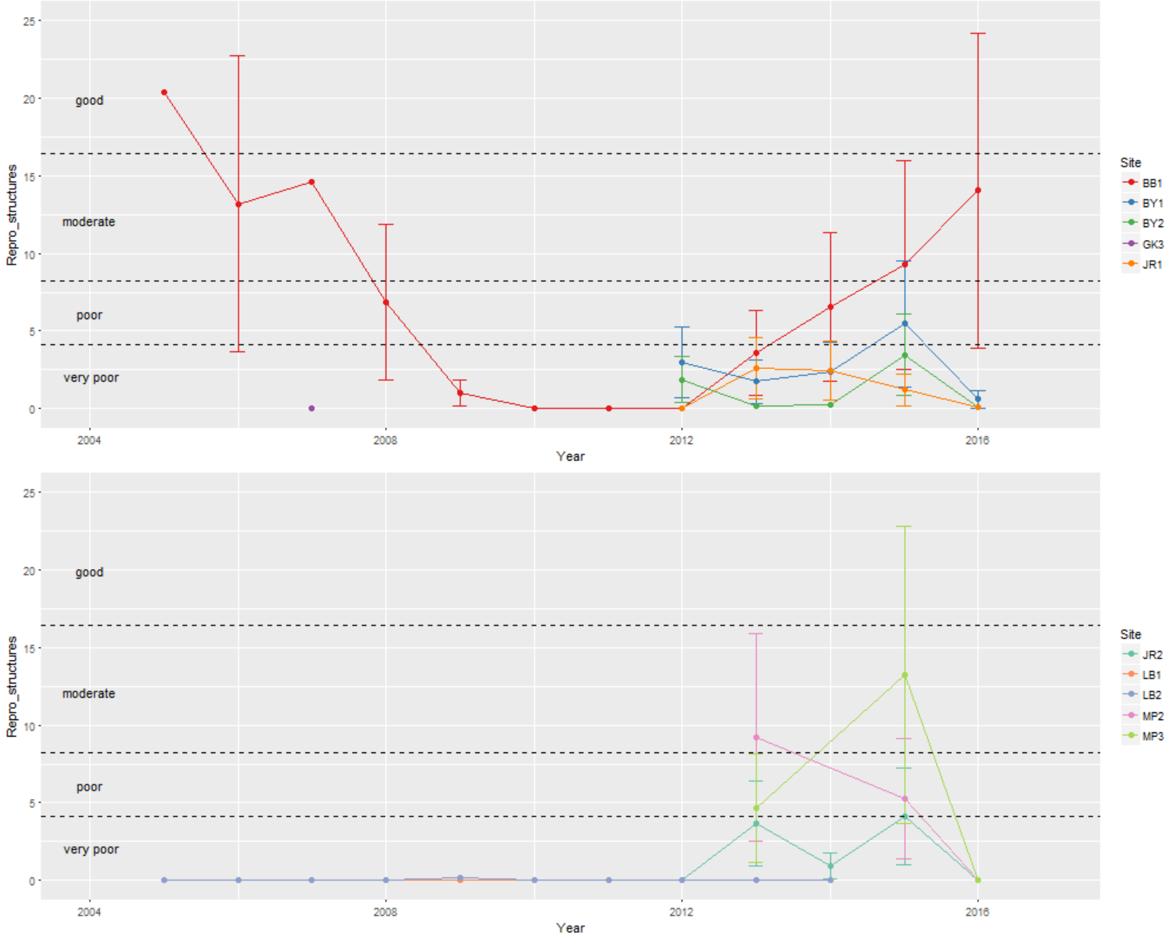

Figure 14: Reproductive effort metric and estimated confidence interval based on negative binomial model for Coastal sites 1 to 10

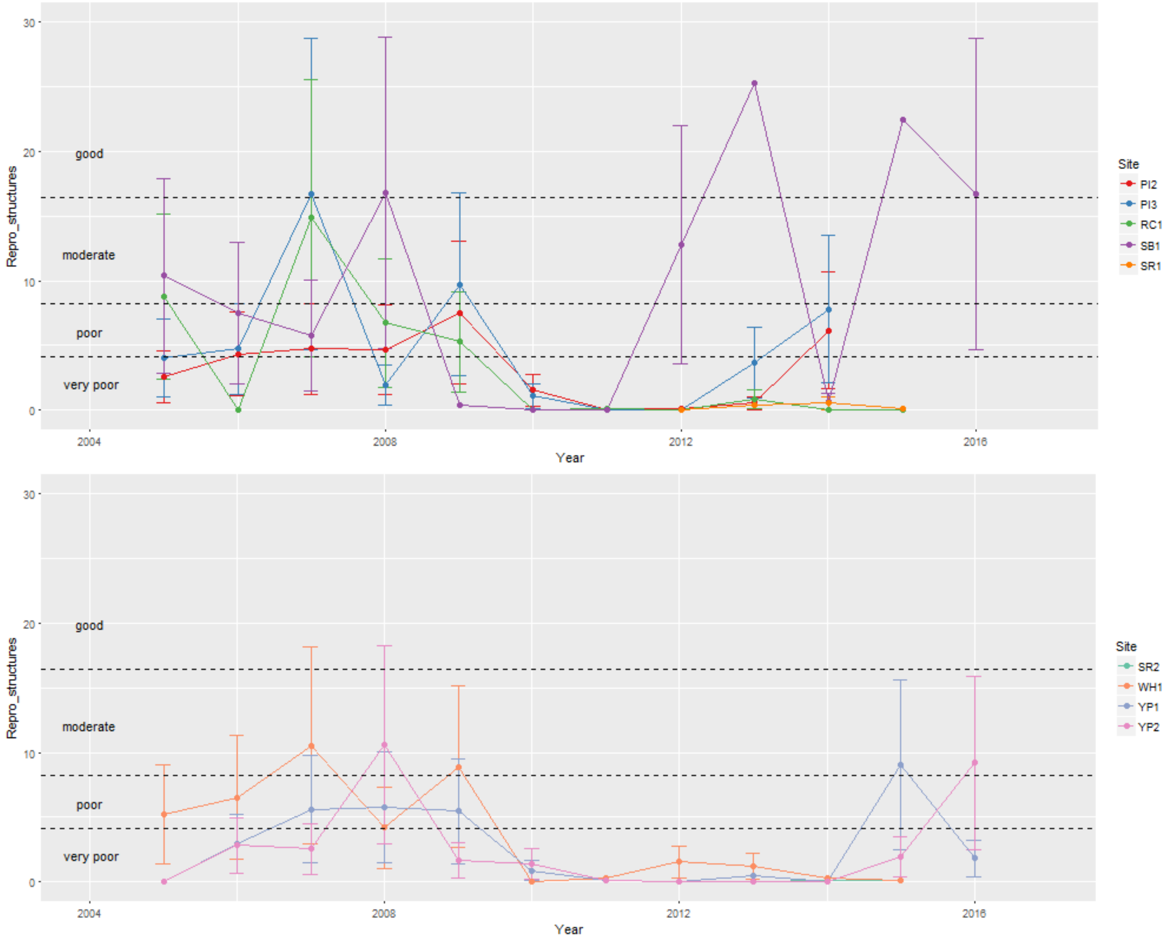

Figure 15: Reproductive effort metric and estimated confidence interval based on negative binomial model for Coastal sites 11 to 19

Figure 16: Reproductive effort metric and estimated confidence interval based on negative binomial model for estuarine sites

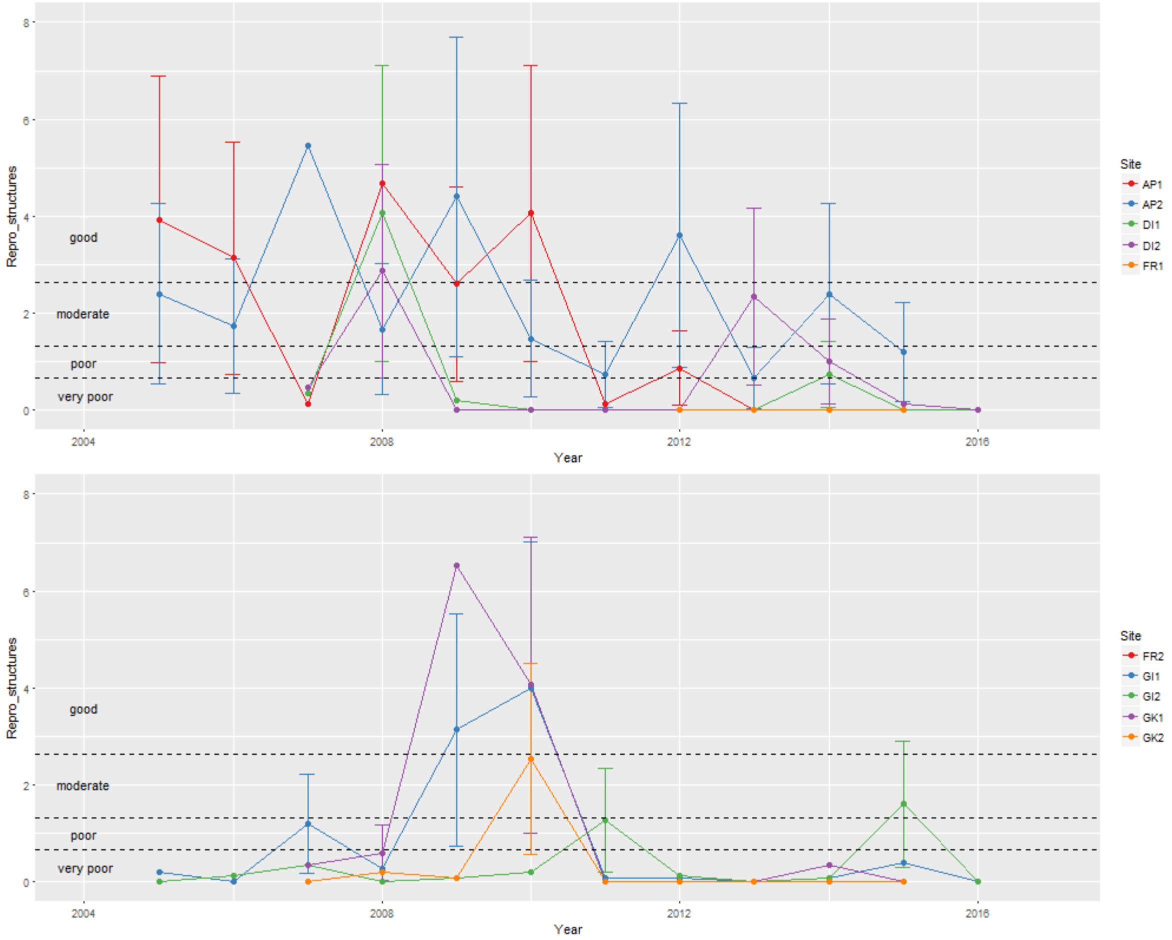

Figure 17: Reproductive effort metric and estimated confidence interval based on negative binomial model for reef sites 1 to 10

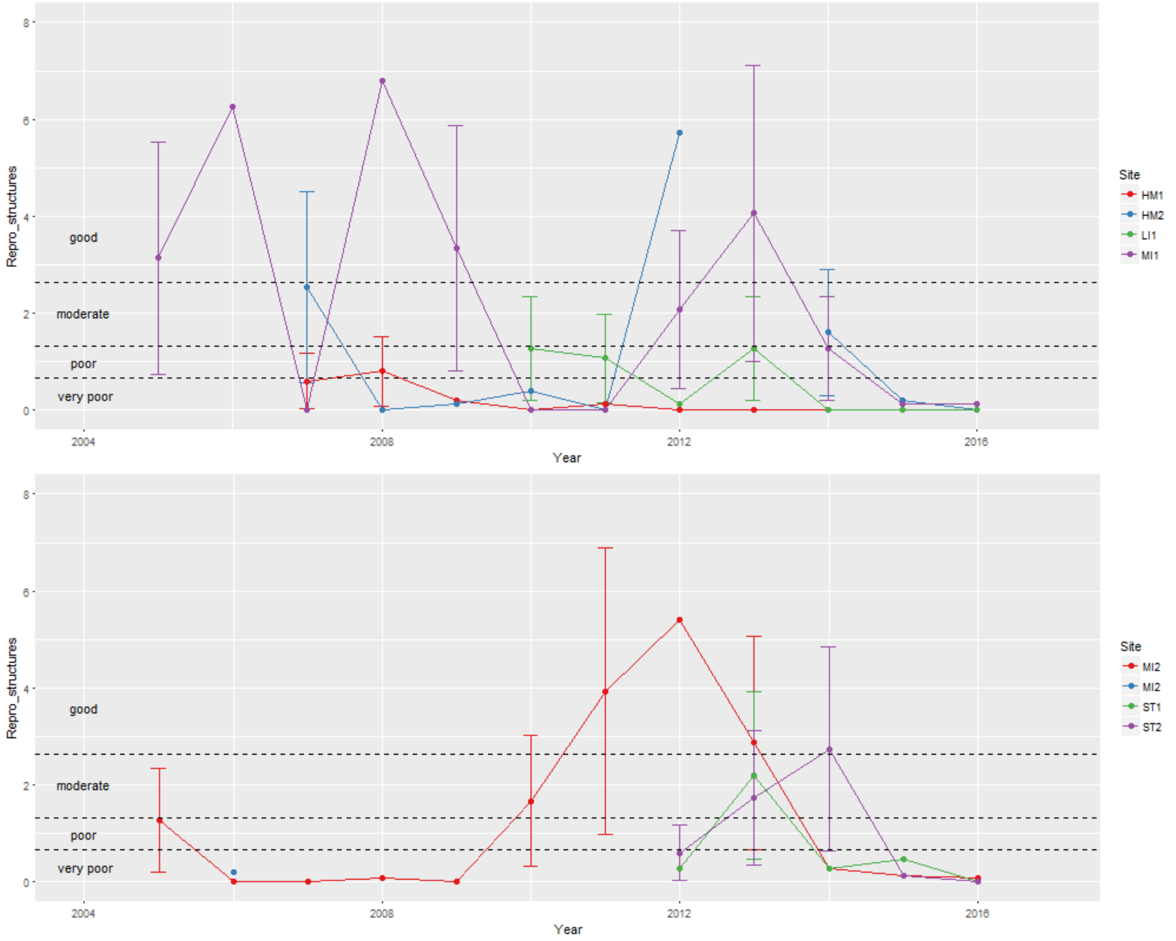

Figure 18: Reproductive effort metric and estimated confidence interval based on negative binomial model for reef sites 11 to 19

The uncertainty is particularly apparent at the reef sites (Figure 17 and Figure 18) where many of the sites span a large range of the report card scores. The scale of the y-axis here is smaller than that of the estuarine and coastal sites so although the confidence intervals are tighter, they appear larger due to the categories being associated with lower-cutoffs in the reef. It would be pertinent to re-visit whether these categories are perhaps a bit ‘tight’ for the reef sites. This is not a statistical question but one of ecological relevancy.

We have effectively treated each site as a location, removing the benefit of replicating by monitoring two sites at a given location. However, we have done this to demonstrate the variability and uncertainty even within some locations. To improve the uncertainty around the reproductive effort measure at each site/location, the power of the metric needs to be improved. The two best ways to do this would be to either take more samples or conduct further investigations around the best time of year at each site to maximise the count of reproductive structures present. We acknowledge that both of these options are potentially costly and difficult due to logistical constraints, but should be considered with regard to management priorities and acceptable levels of uncertainty.

Another option to significantly improve the power to detect change in reproductive effort at a broader spatial scale would be to move to a design-based sampling regime as described in Udy et al. 2019. This would significantly improve the inference at the habitat and region level by allowing the trends from sites to be aggregated in a statistically unbiased manner. This is, however, a substantial change.

References

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J. 1984, Classification and Regression Trees. Wadsworth, Monterey, CA, USA.

Collier, C., and M. Waycott. 2009. Drivers of change to seagrass distributions and communities on the Great Barrier Reef: Literature review and gaps analysis. Reef and Rainforest Research Centre Limited, Cairns.

De'ath, G. 2002, Multivariate Regression Trees: A new technique for modelling species-environment relationships. Ecology 83 (4), 1105-1117.

Kuhnert, P.M., Liu, Y., Henderson, B.L., Dambacher, J., Lawrence, E., and Kroon, F.J. 2015, Review of the Marine Monitoring Program (MMP), Final Report for the Great Barrier Reef Marine Park Authority (GBRMPA). Adelaide: CSIRO.

Liaw A. and Wiener M. 2002, Classification and Regression by randomForest. *R News*. 2(3):18-22.

McKenzie, L.J., Collier, C.J, Langlois, L.A, Yoshida, R.L, Smith, N. and Waycott, M. 2016, Marine Monitoring Program. Annual Report for inshore seagrass monitoring: 2014 to 2015. Report for the Great Barrier Reef Marine Park Authority. TropWATER, James Cook University, Cairns. 236pp.

Nelder J. and Wedderburn R.W M. 1972, Generalized Linear Models, *J. R. Statist. Soc.* A, 135: 370–384.

Ospina R. and Ferrari S.L.P. 2010, Inflated beta distributions, *Statistical Papers*, 23:11-126

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rigby R.A. and Stasinopoulos D.M. 2005, Generalized additive models for location, scale and shape, *Applied Statistics,* 54(3):507-554.

Stasinopoulos D. M., Rigby R.A., Heller G., Voudouris V., and De Bastiani F. 2017, Flexible Regression and Smoothing: Using GAMLSS in R, Chapman and Hall/CR

Udy, J., et al. 2019, Monitoring seagrass within the Reef 2050 Integrated Monitoring and Reporting Program: Final report of the seagrass expert group. Report for the Great Barrier Reef Marine Park Authority, Great Barrier Reef Marine Park Authority, Townsville (In press).

Unsworth, R.K.F., Collier, C.J., Waycott, M., McKenzie, L.J., and Cullen-Unsworth, L.C. 2015, A framework for the resilience of seagrass ecosystems. *Marine Pollution Bulletin, 100*(1), 34-46. doi:http://dx.doi.org/10.1016/j.marpolbul.2015.08.016

Appendix A. Summary of reproductive effort data for each site

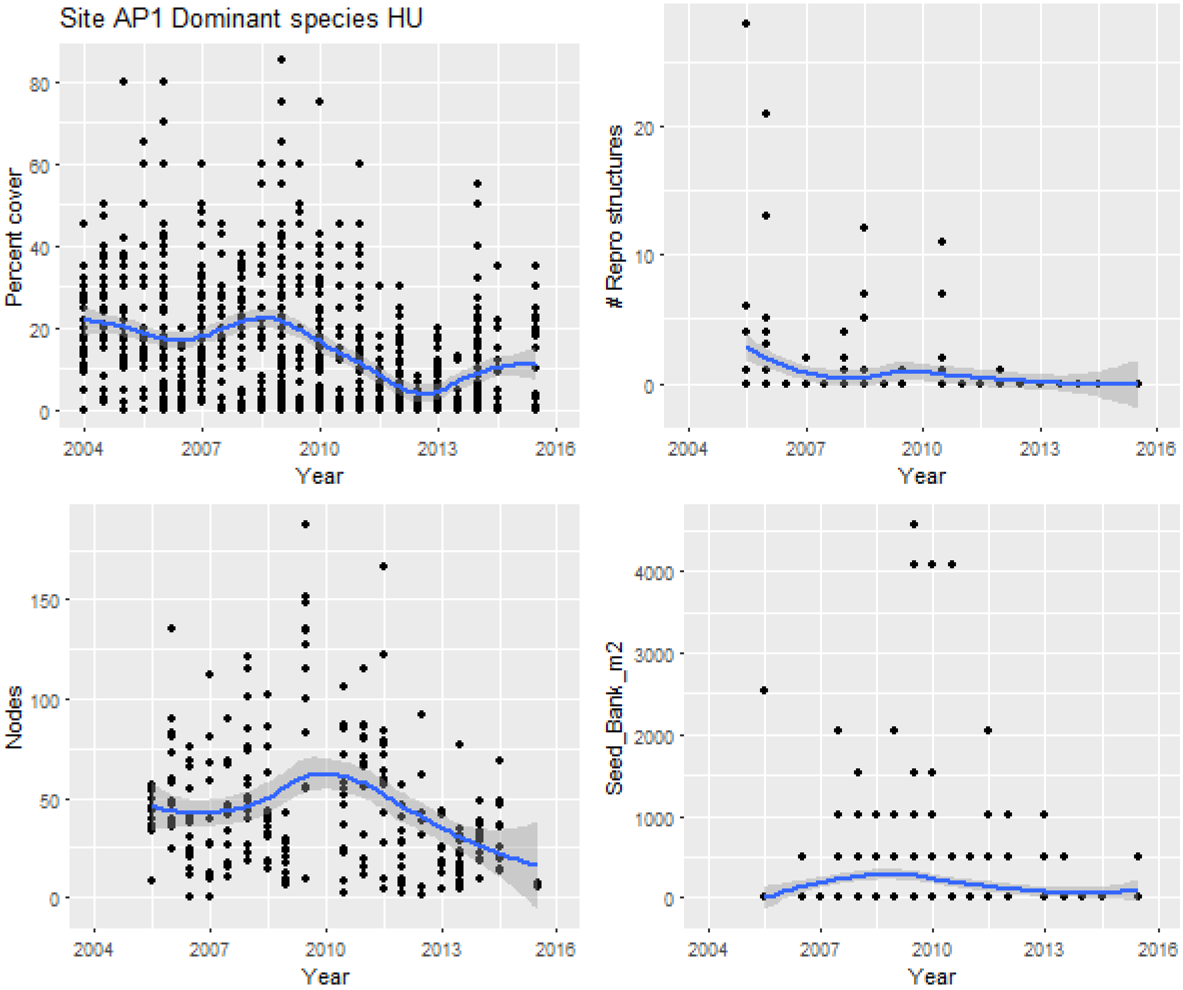

Figure A1: Summary of reproductive effort data for Archer Point 1, dominant species Halodule.

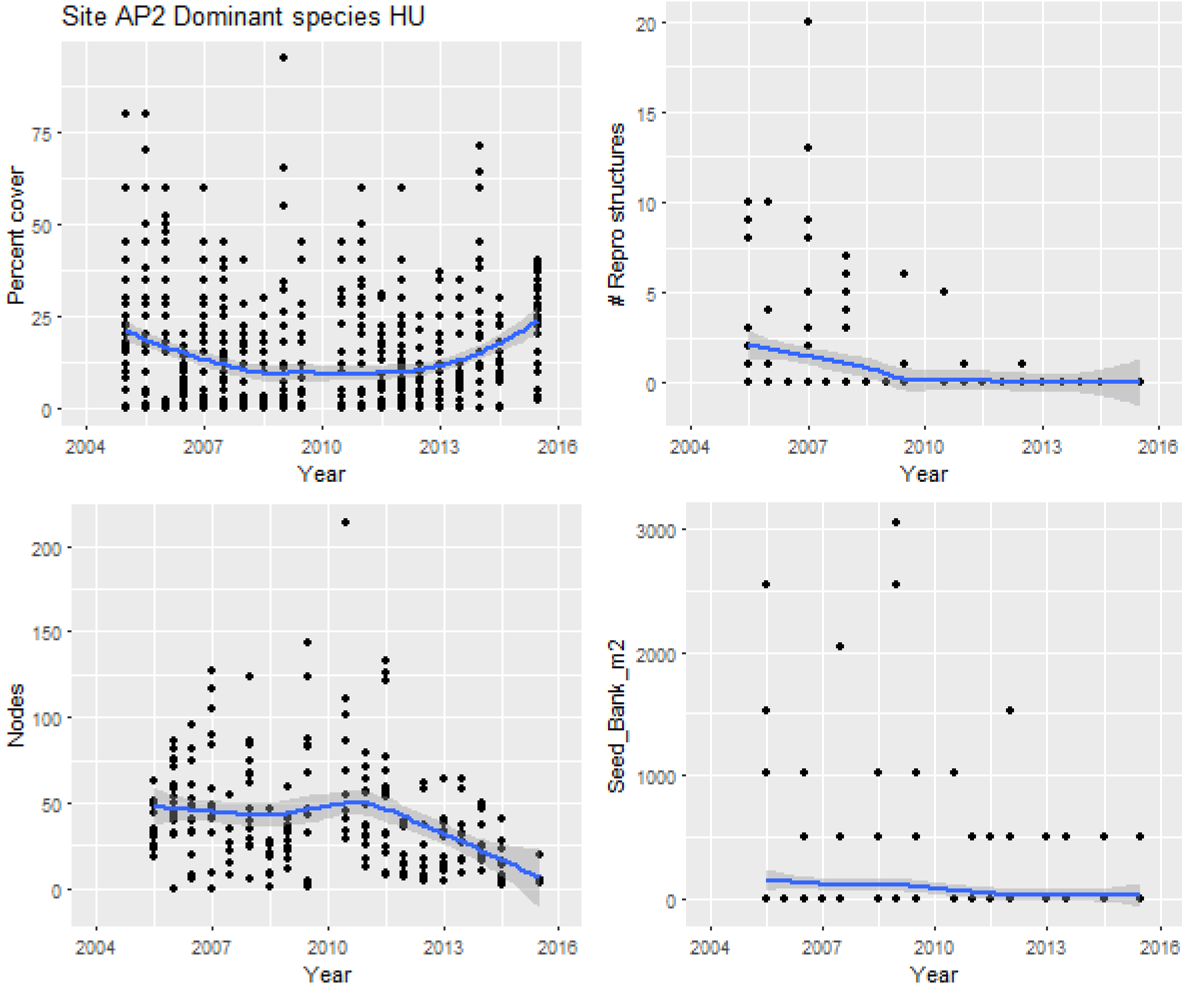

Figure A2: Summary of reproductive effort data for Archer Point 2, dominant species Halodule.

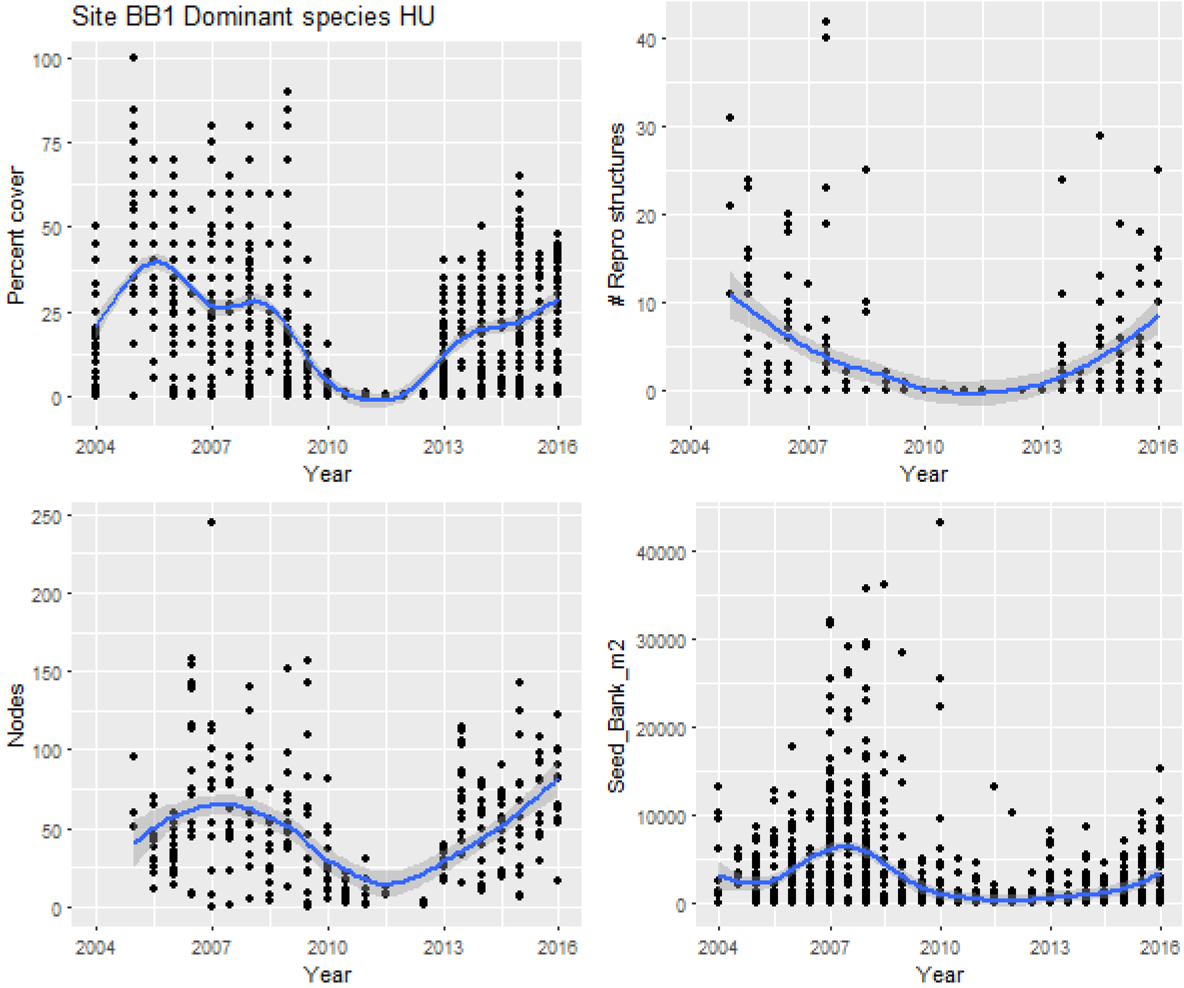

Figure A3: Summary of reproductive effort data for Bushland Beach, dominant species Halodule.

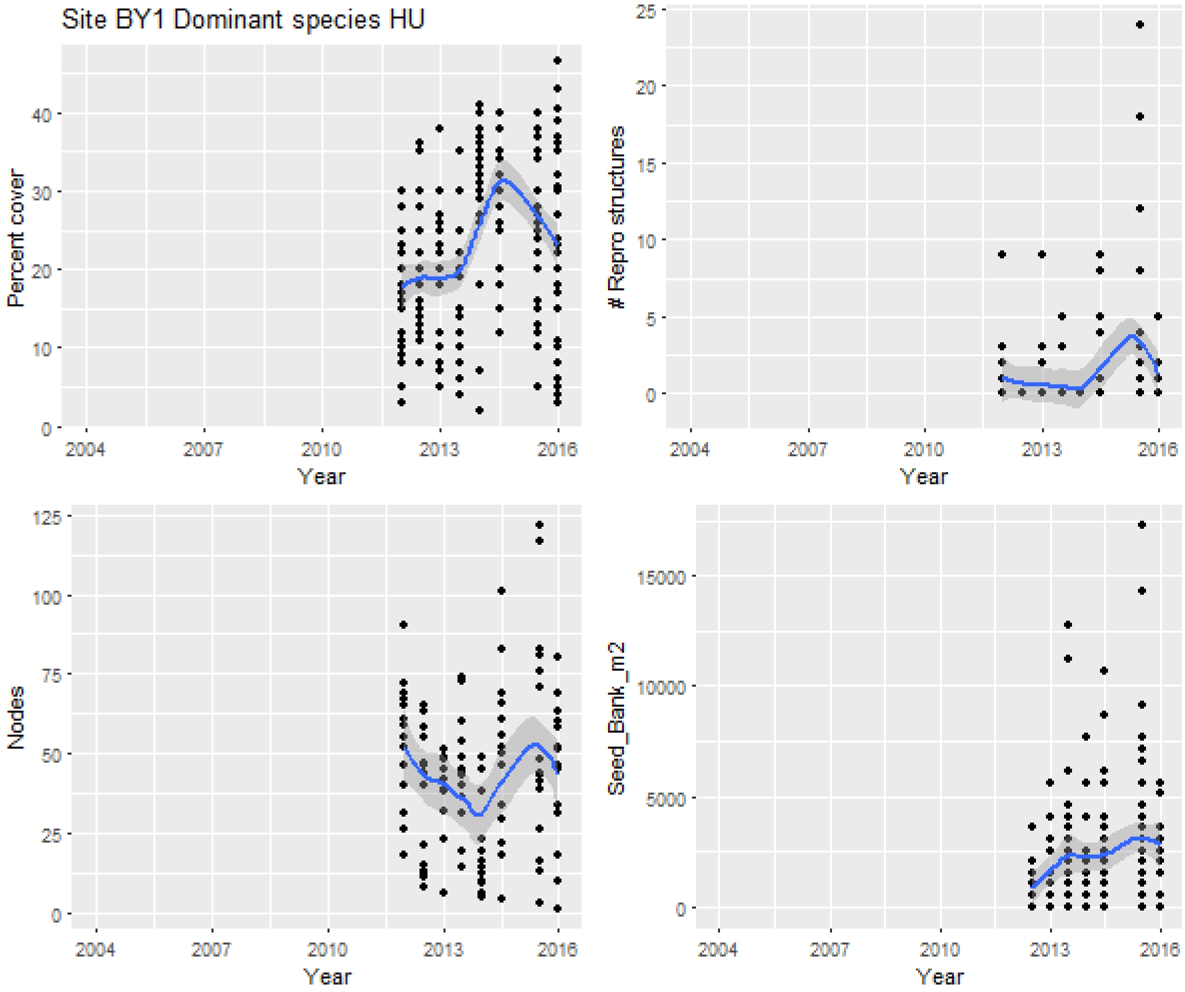

Figure A4: Summary of reproductive effort data for Bathurst Bay 1, dominant species Halodule.

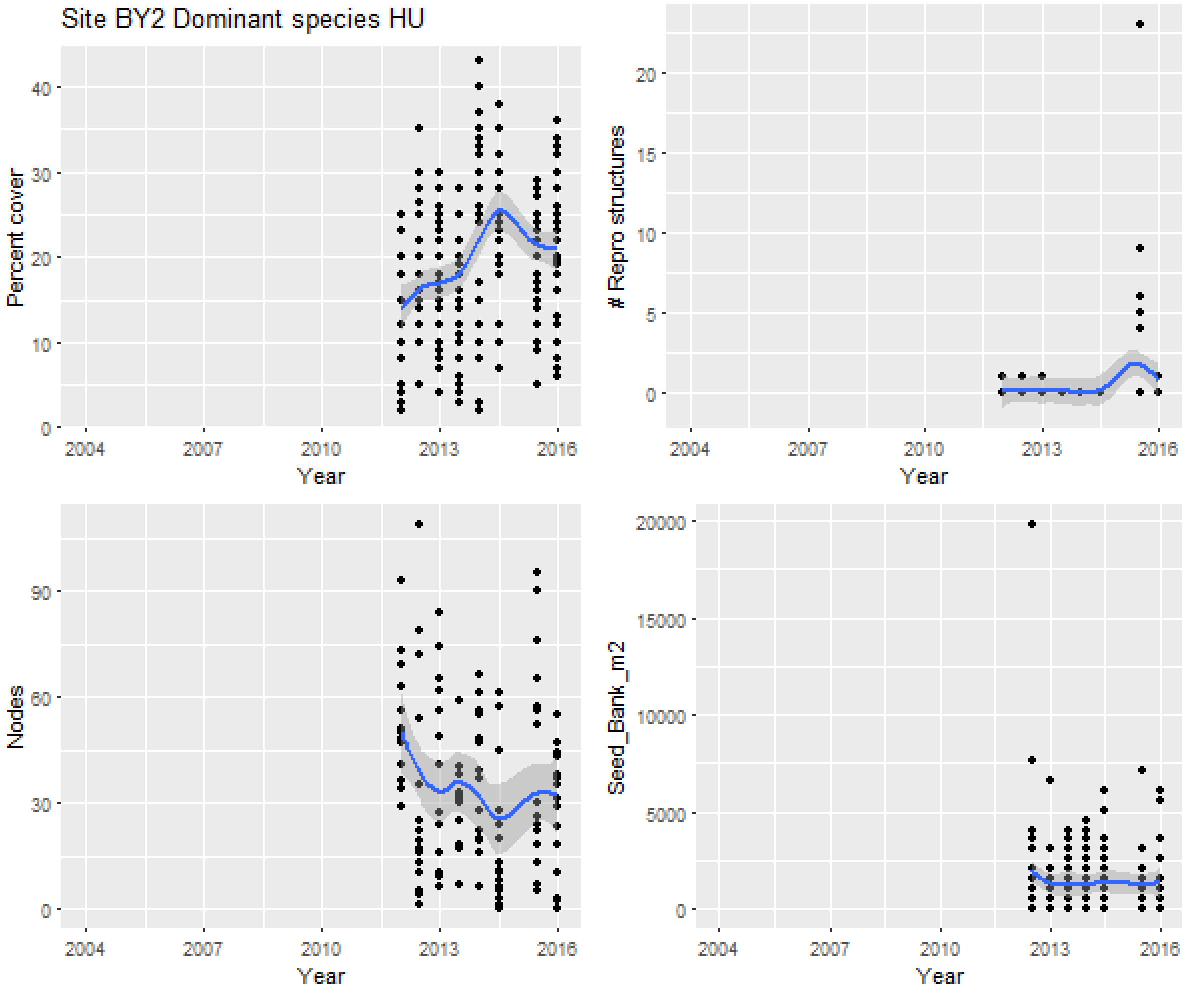

Figure A5: Summary of reproductive effort data for Bathurst Bay 2, dominant species Halodule.

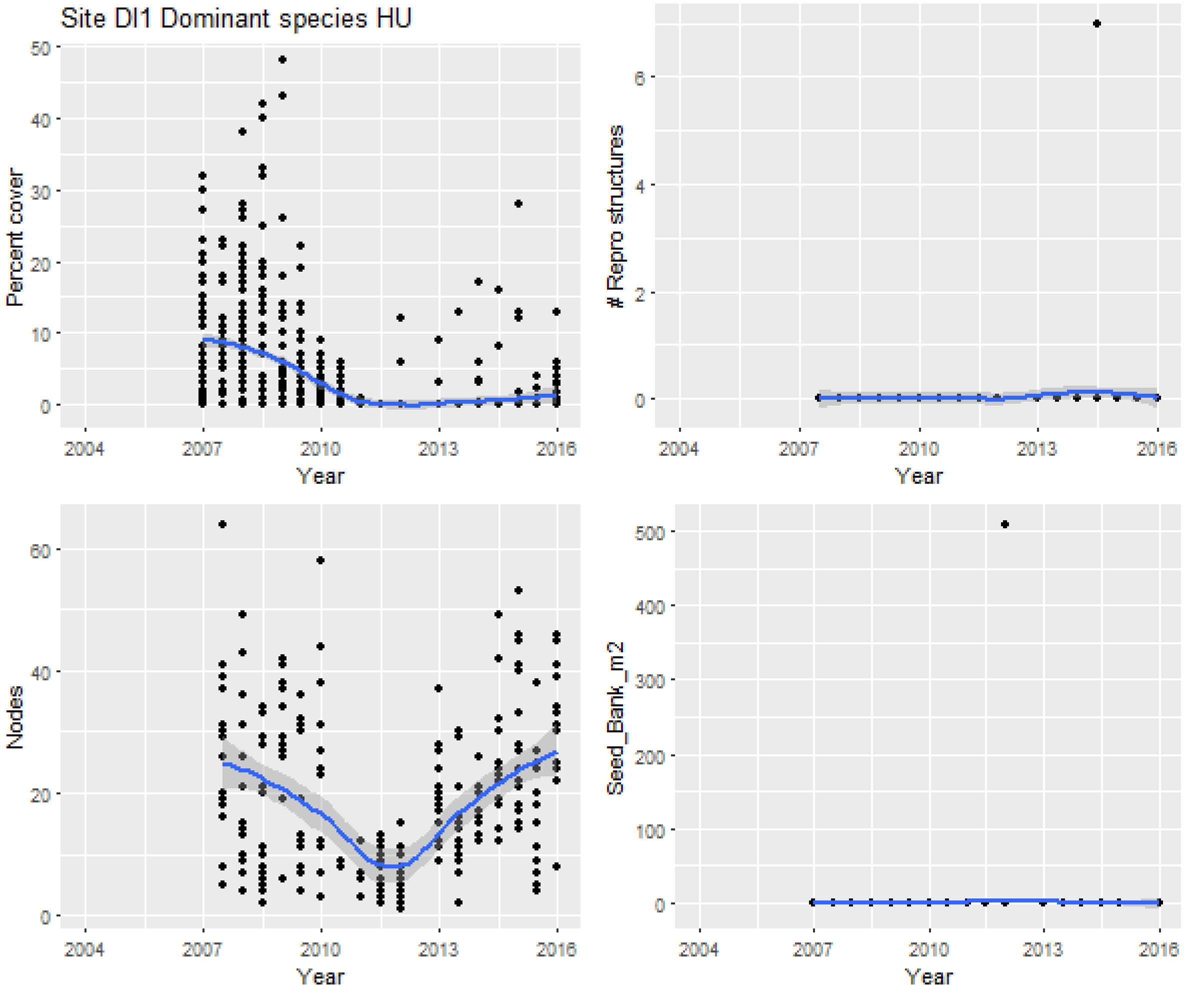

Figure A6: Summary of reproductive effort data for Dunk Island 1, dominant species Halodule.

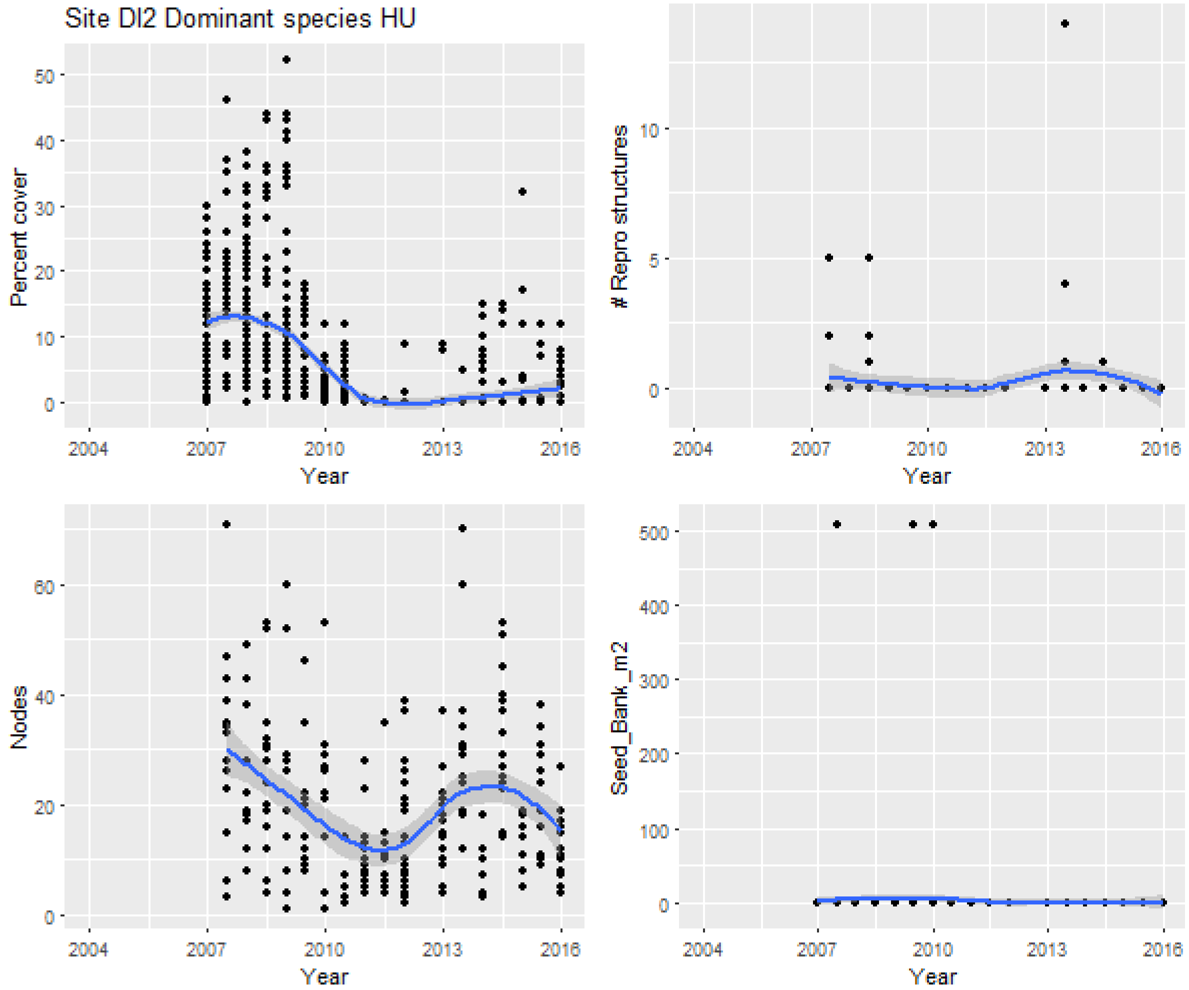

Figure A7: Summary of reproductive effort data for Dunk Island 2, dominant species Halodule.

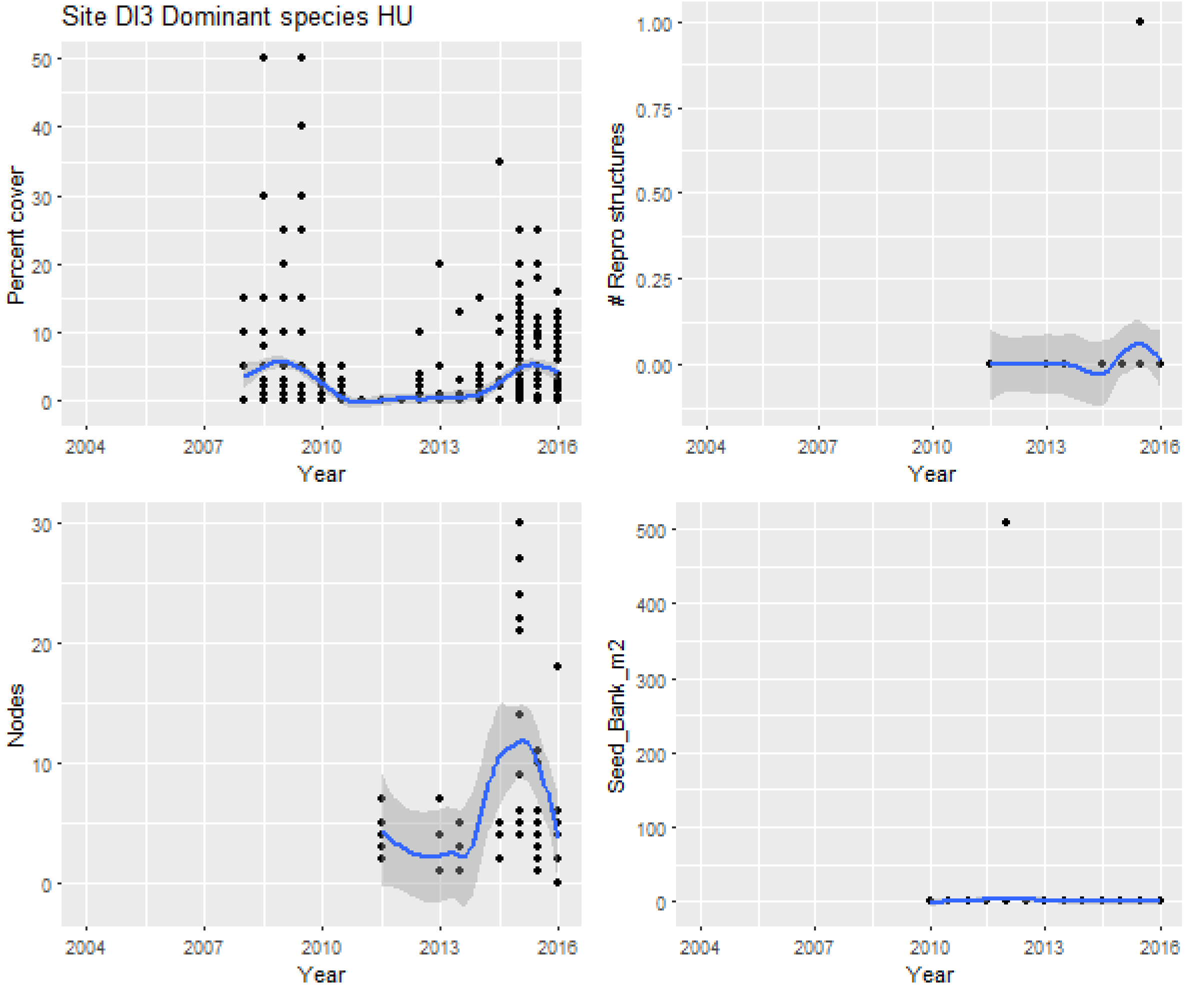

Figure A8: Summary of reproductive effort data for Dunk Island, dominant species Halodule.

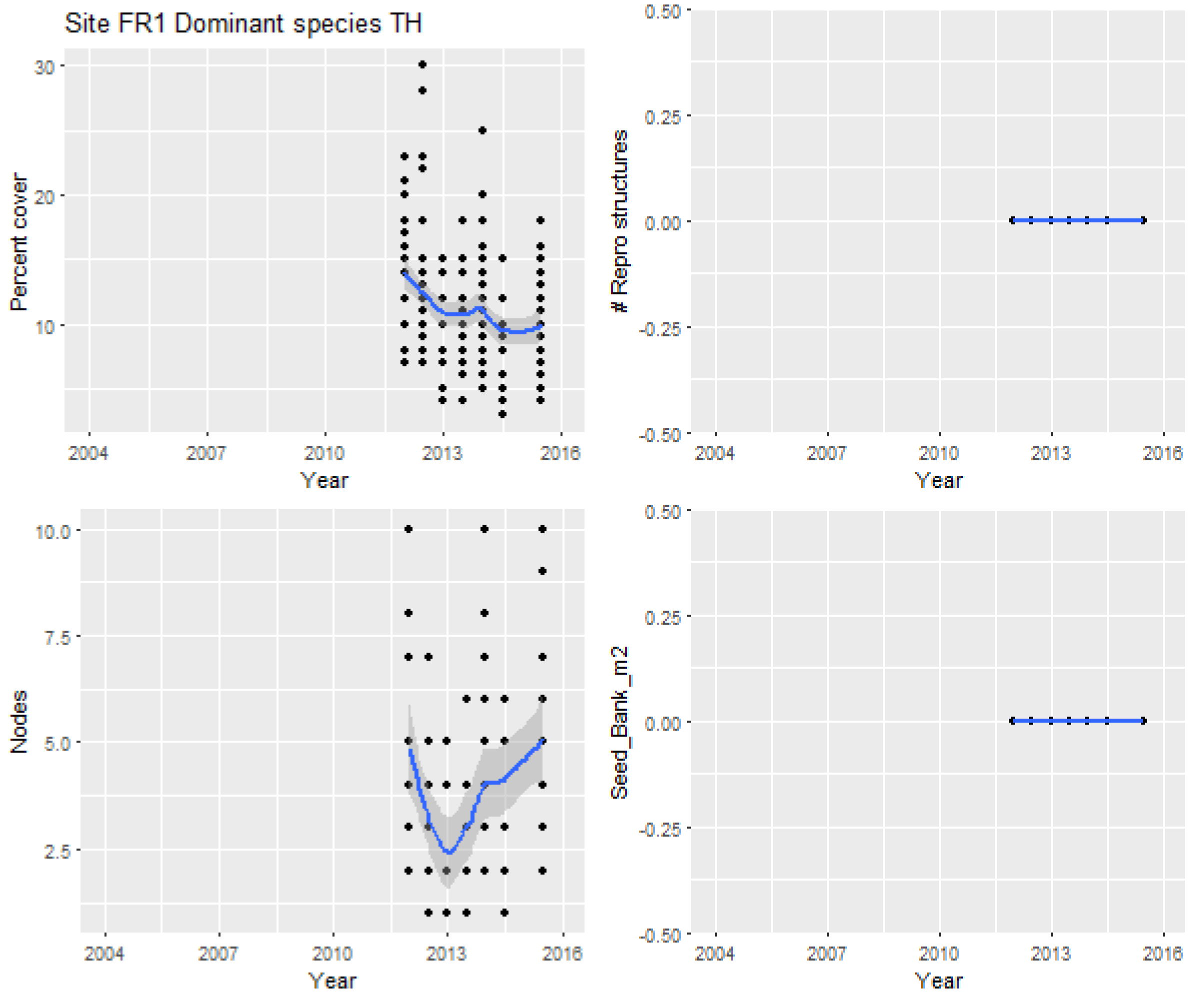

Figure A9: Summary of reproductive effort data for Farmer Is. 1, dominant species Thalassia.

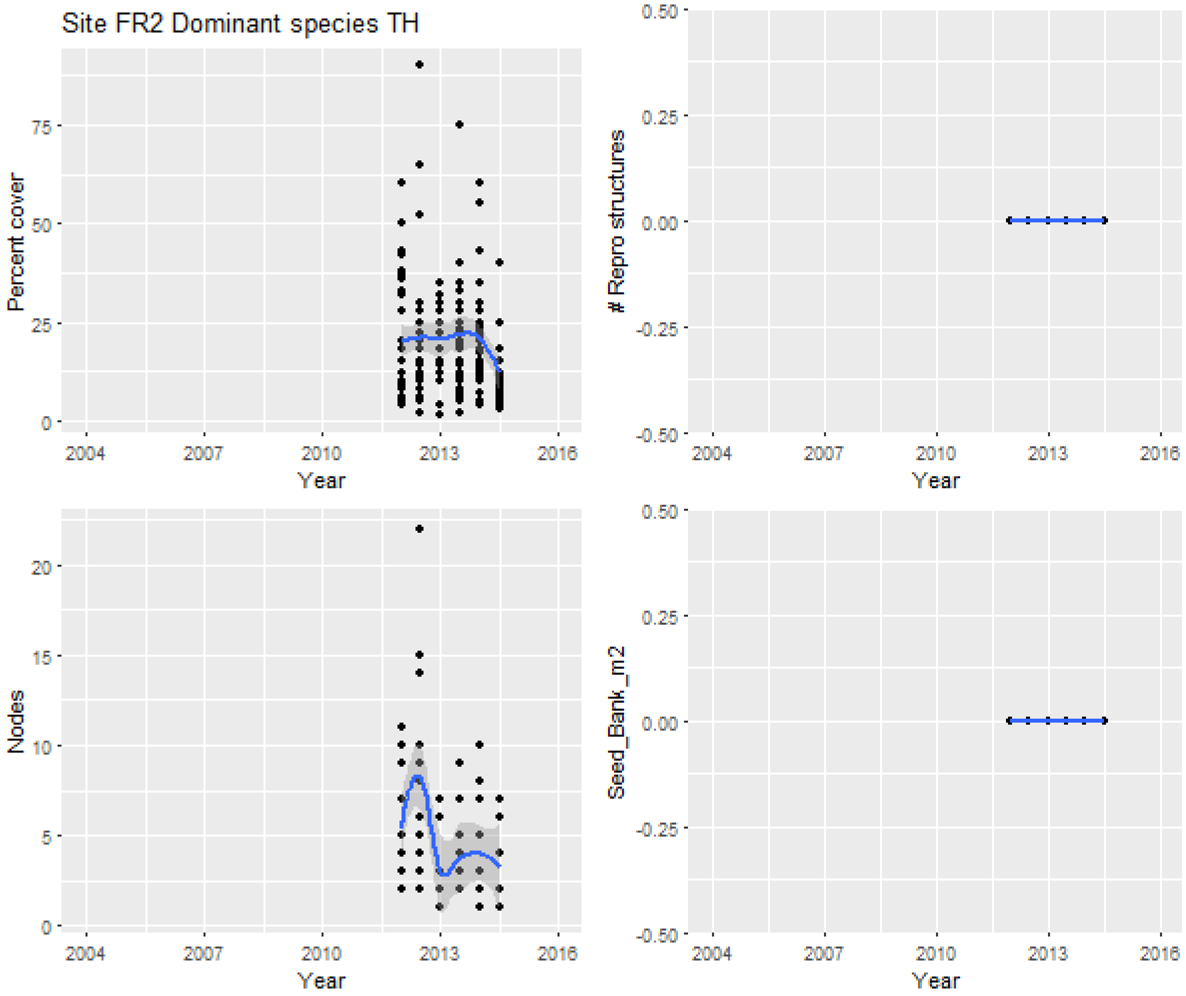

Figure A10: Summary of reproductive effort data for Farmer Is. 2, dominant species Thalassia.

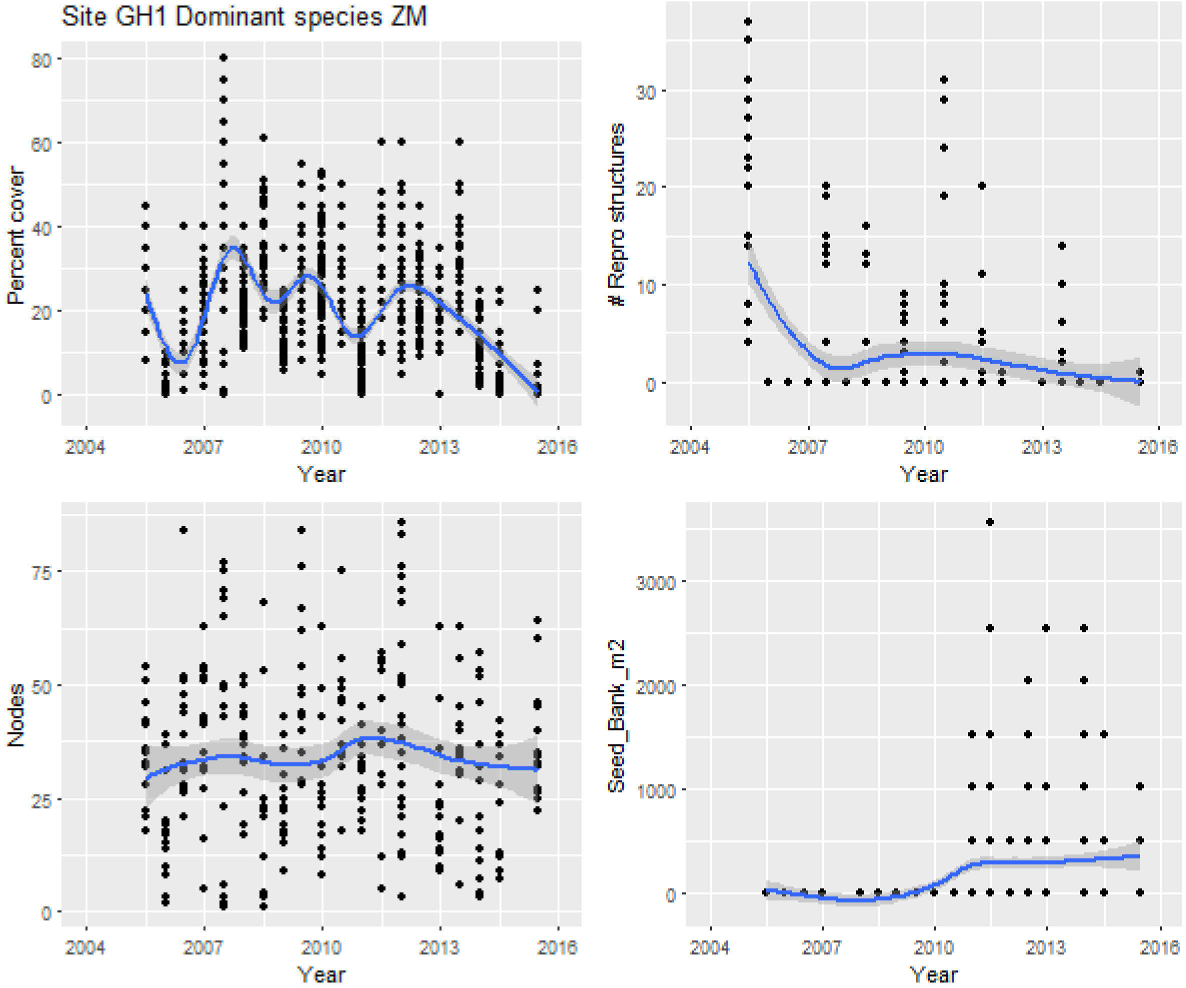

Figure A11: Summary of reproductive effort data for Gladstone Harbour 1, dominant species Zostera.

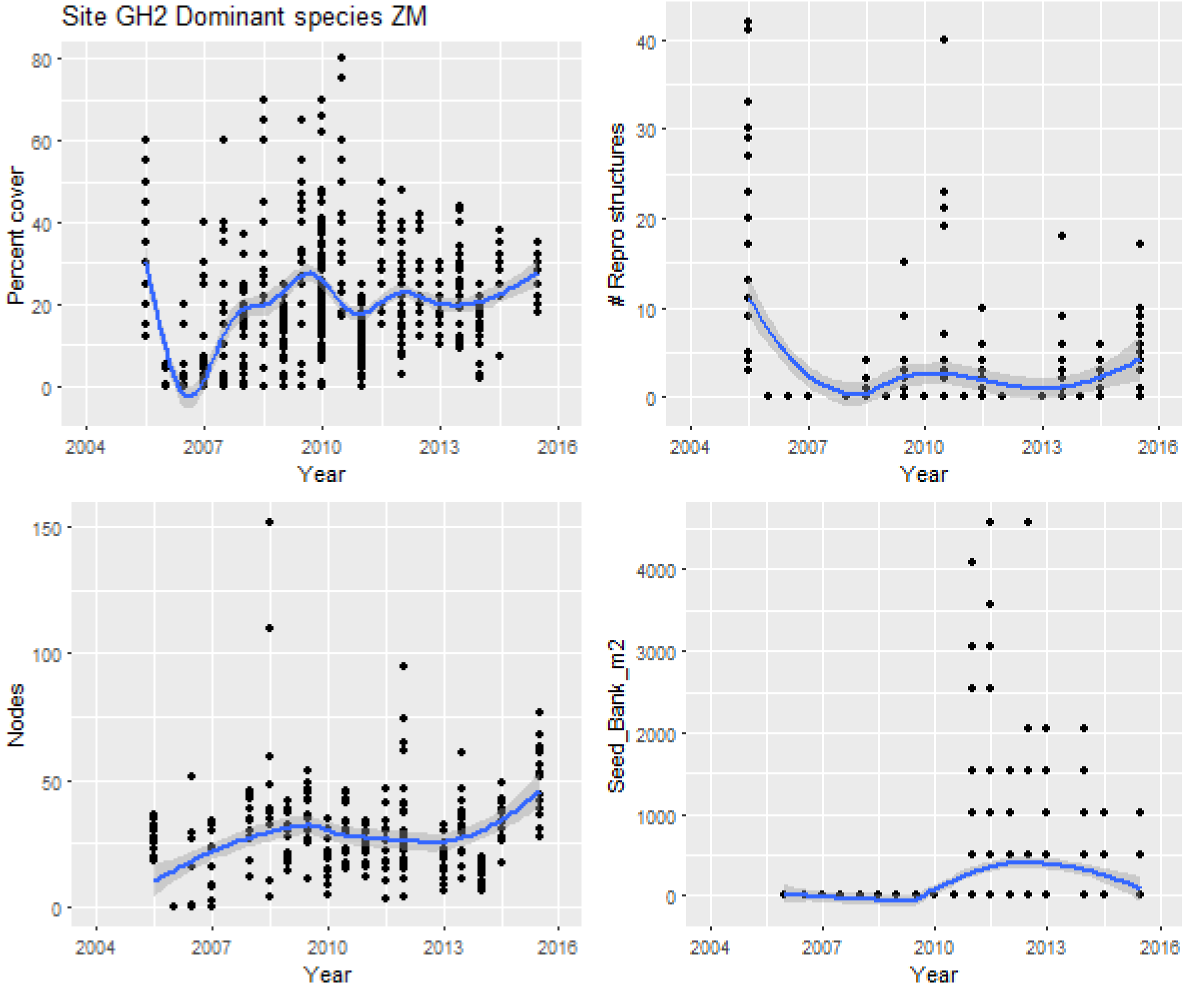

Figure A12: Summary of reproductive effort data for Gladstone Harbour 2, dominant species Zostera..

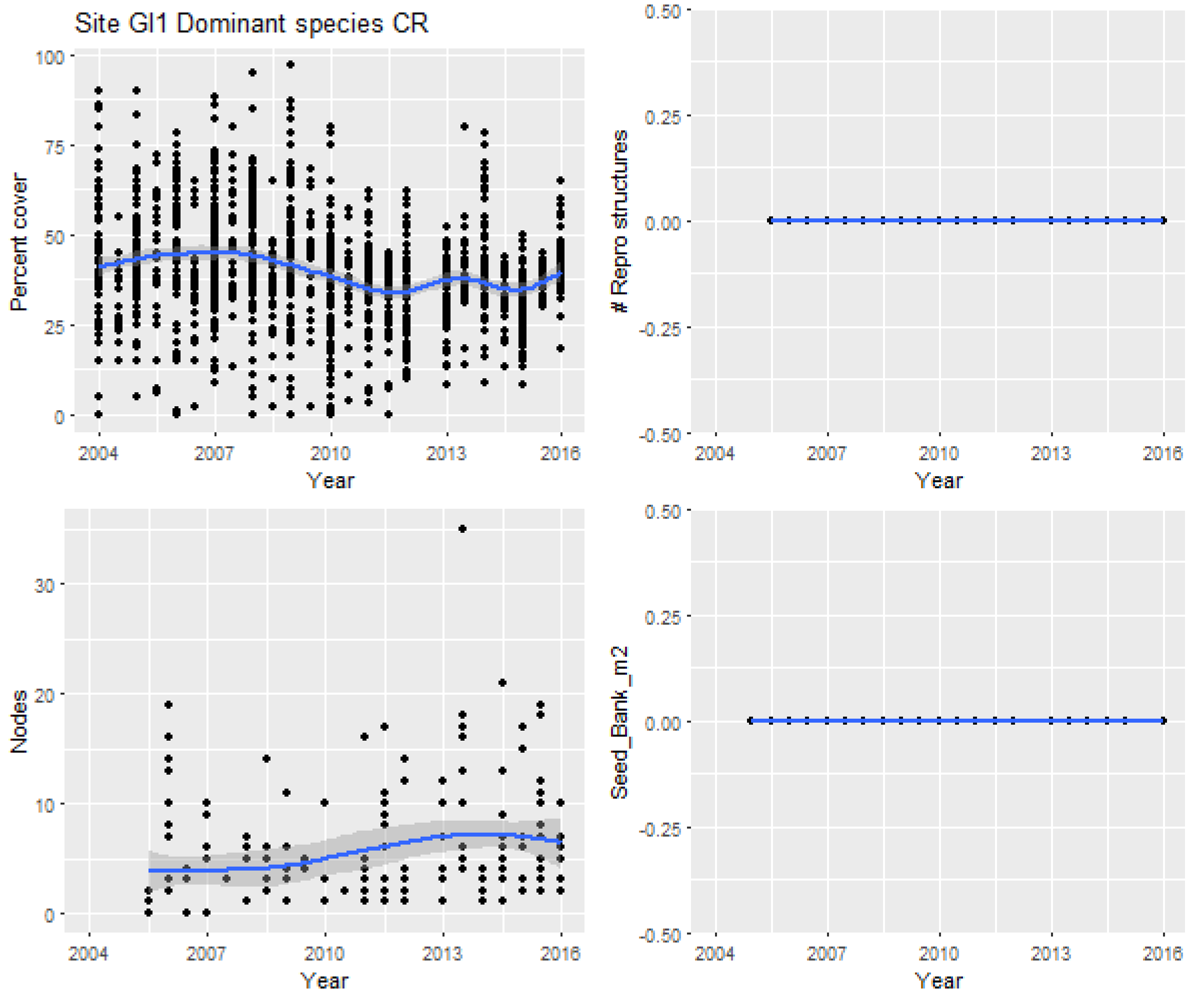

Figure A13: Summary of reproductive effort data for Green Island 1, dominant species Cymodocea rotundata.

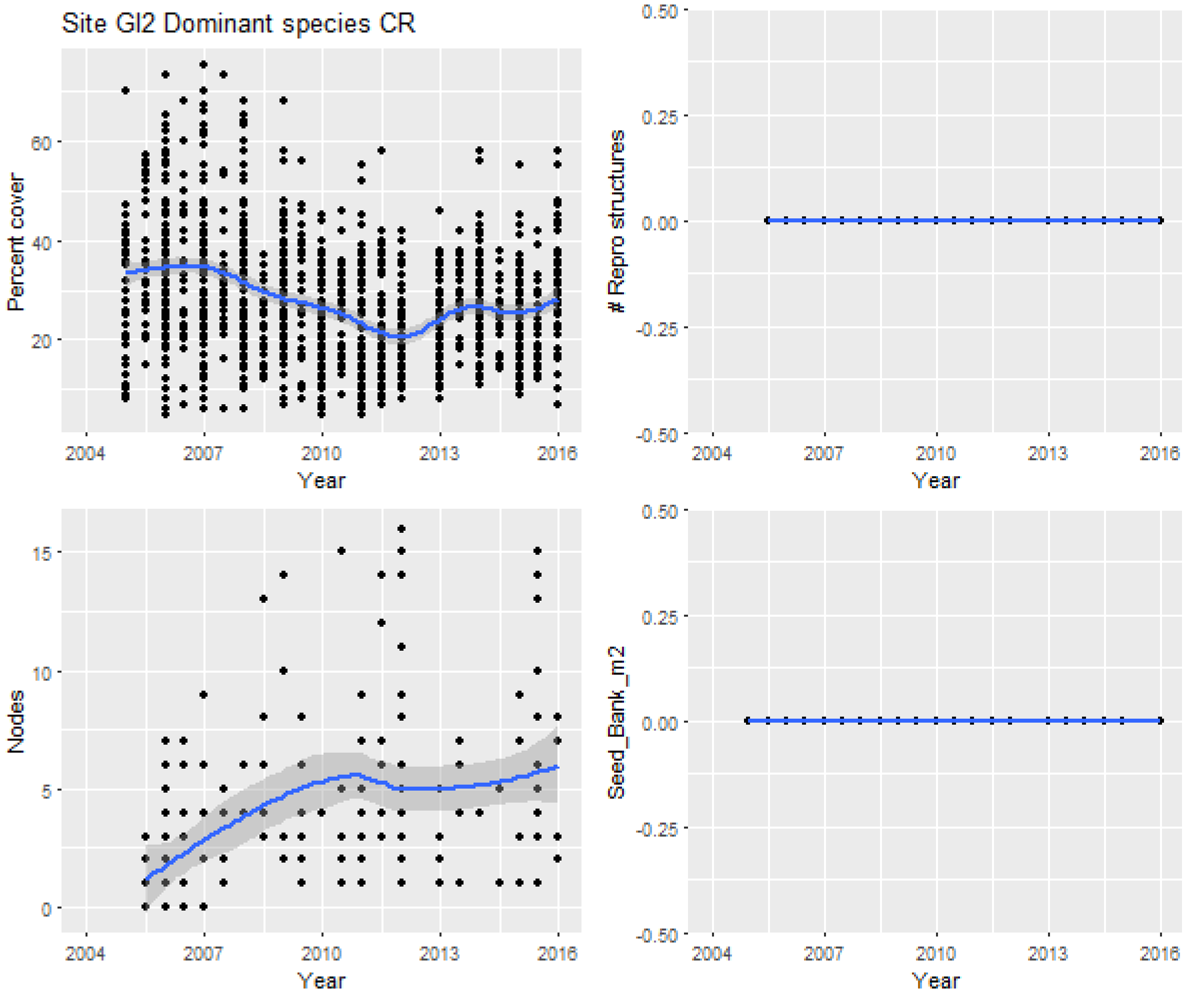

Figure A14: Summary of reproductive effort data for Green Island 2, dominant species Cymodocea rotundata.

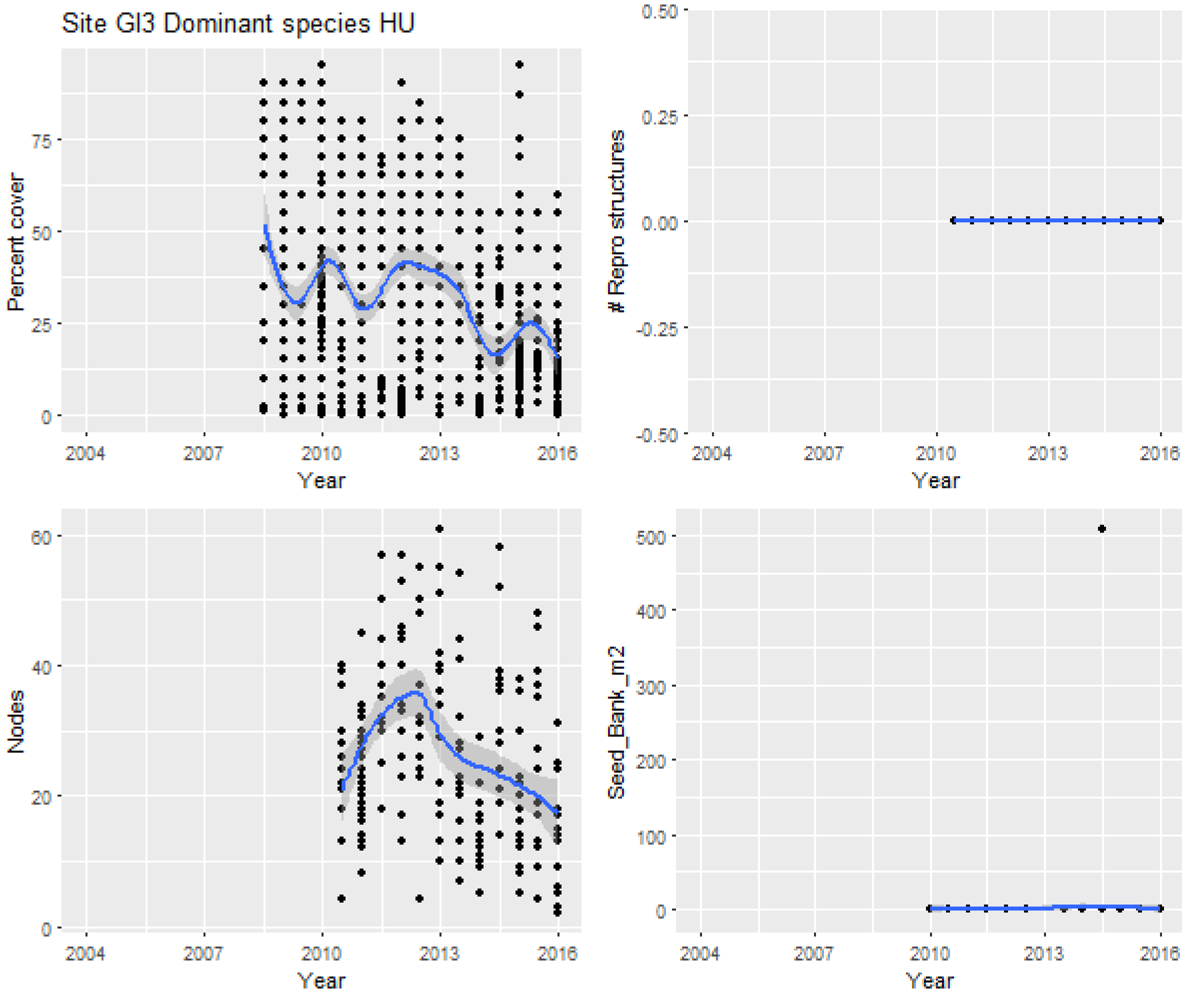

Figure A15: Summary of reproductive effort data for Green Island 3, dominant species Halodule.

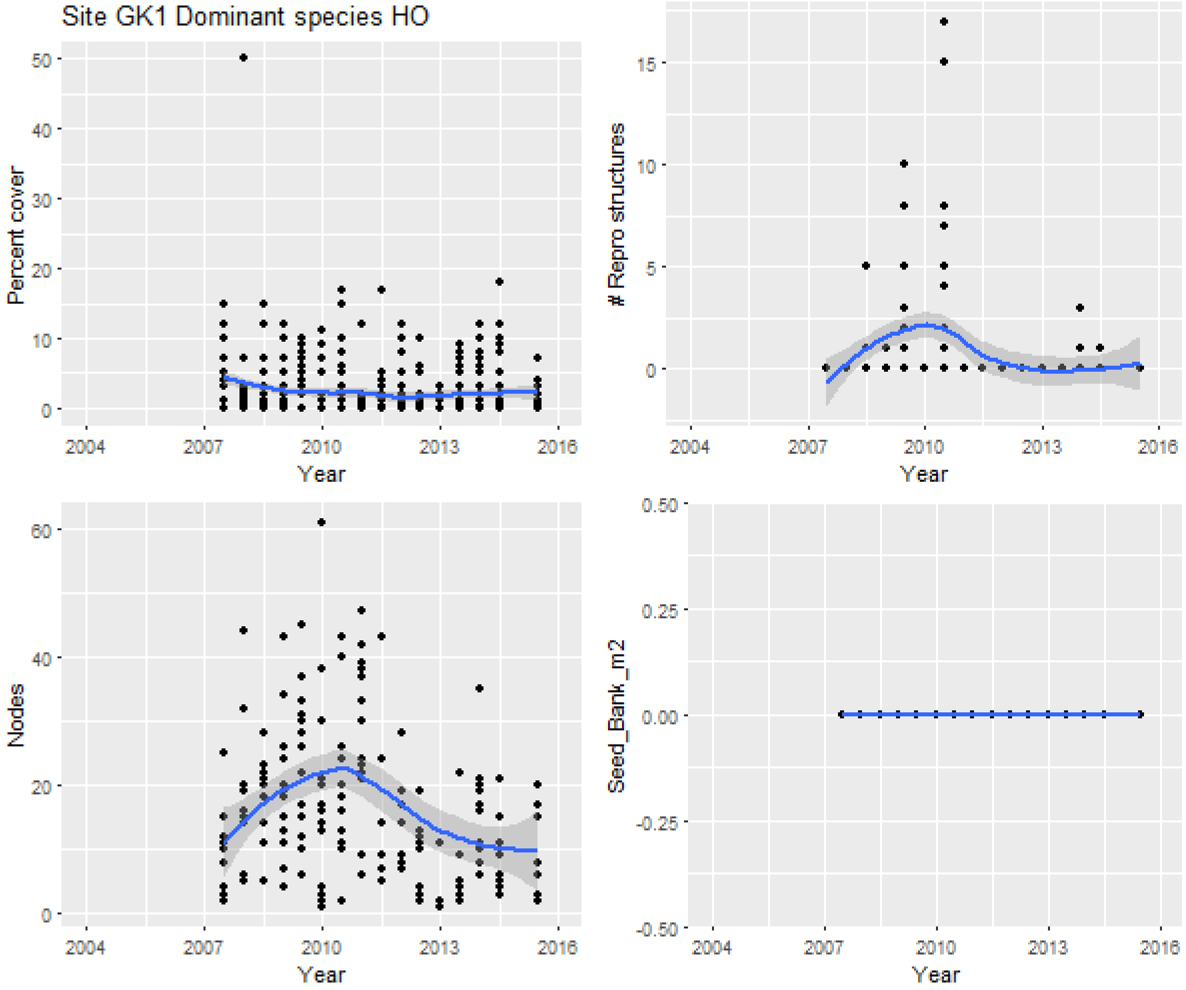

Figure A16: Summary of reproductive effort data for Great Keppel Island 1, dominant species Halophila ovalis.

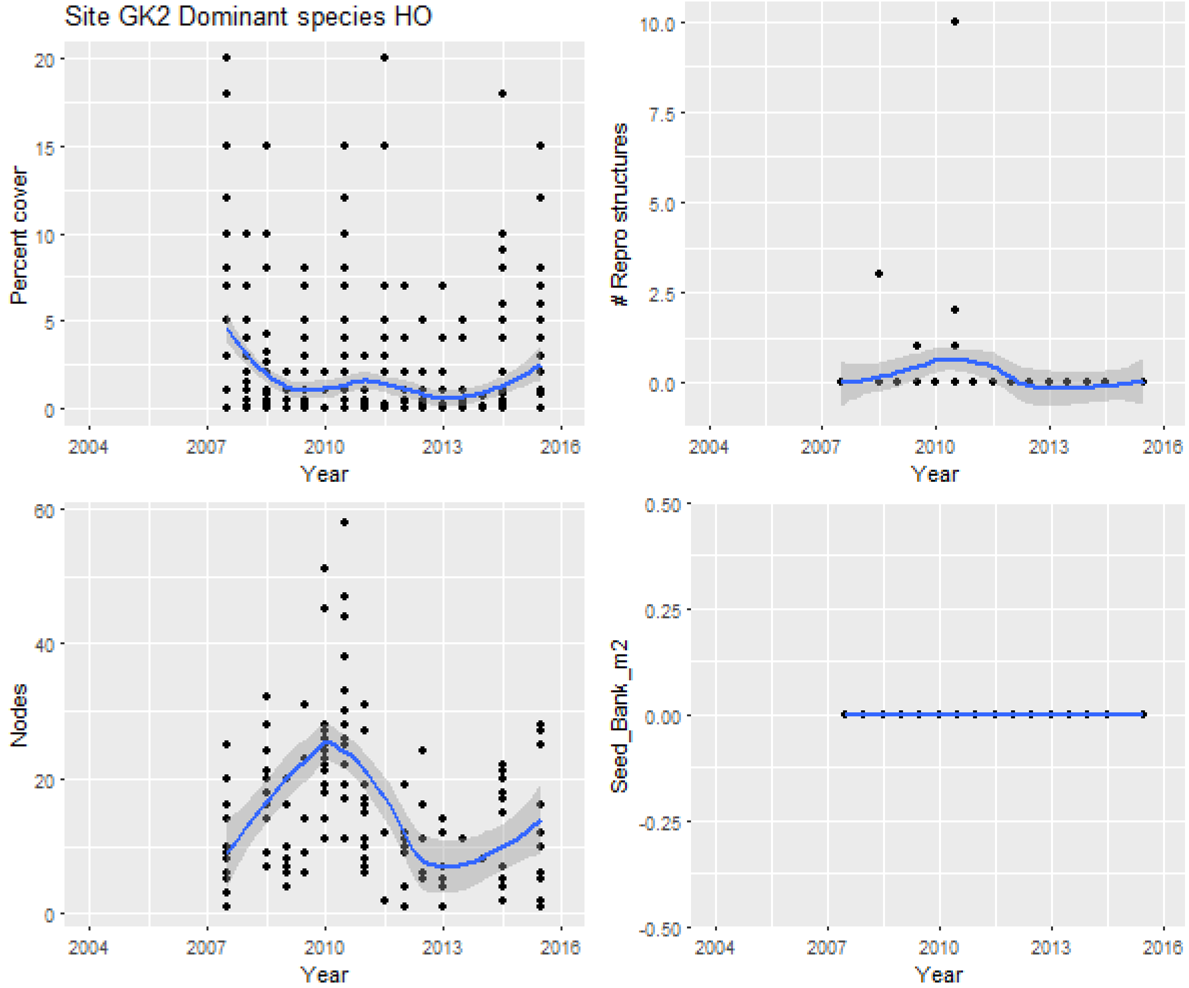

Figure A17: *Summary of reproductive effort data for Great Keppel Island 2, dominant species Halophila ovalis.*

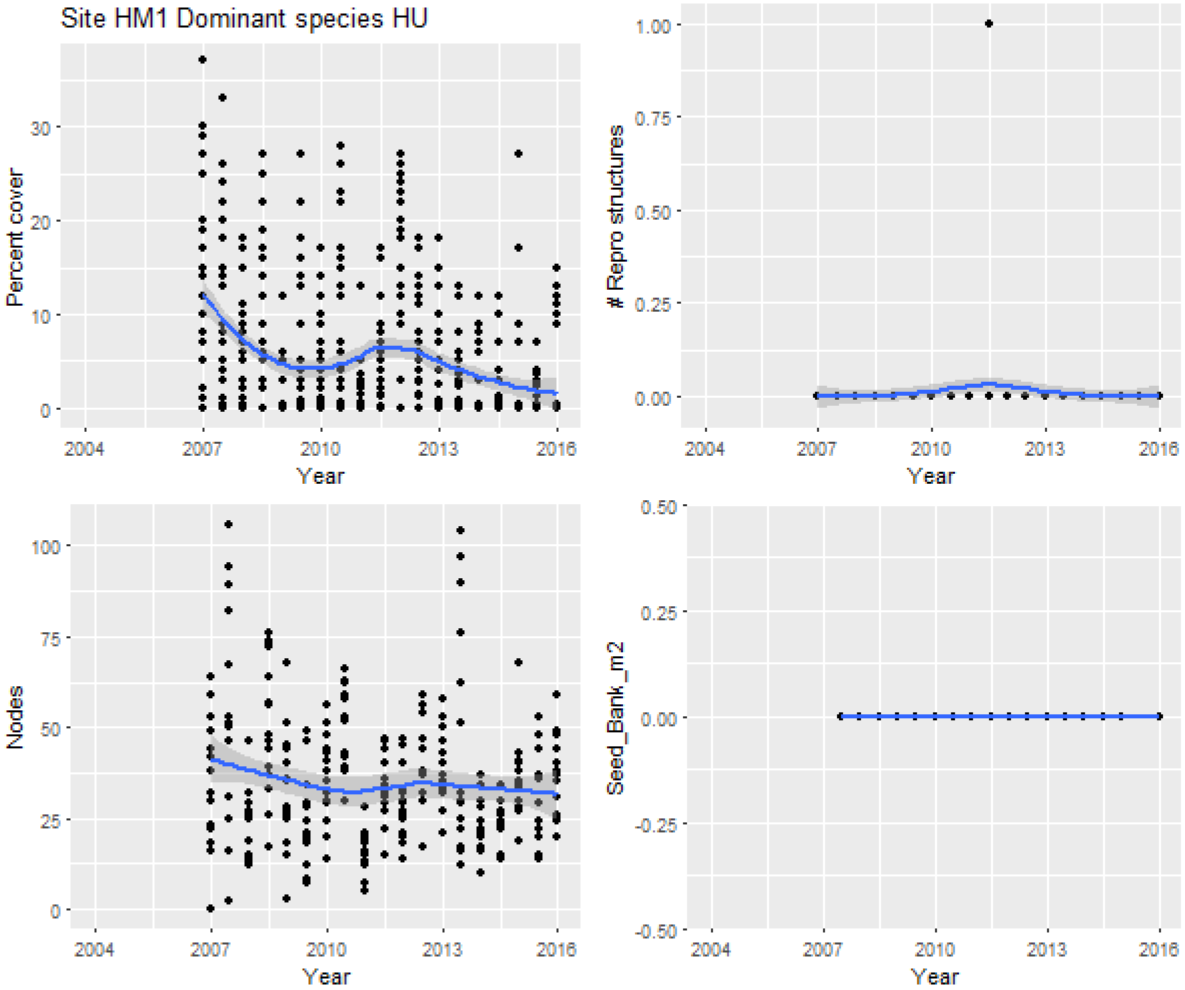

Figure A18: Summary of reproductive effort data for Hamilton Island 1, dominant species Halodule.

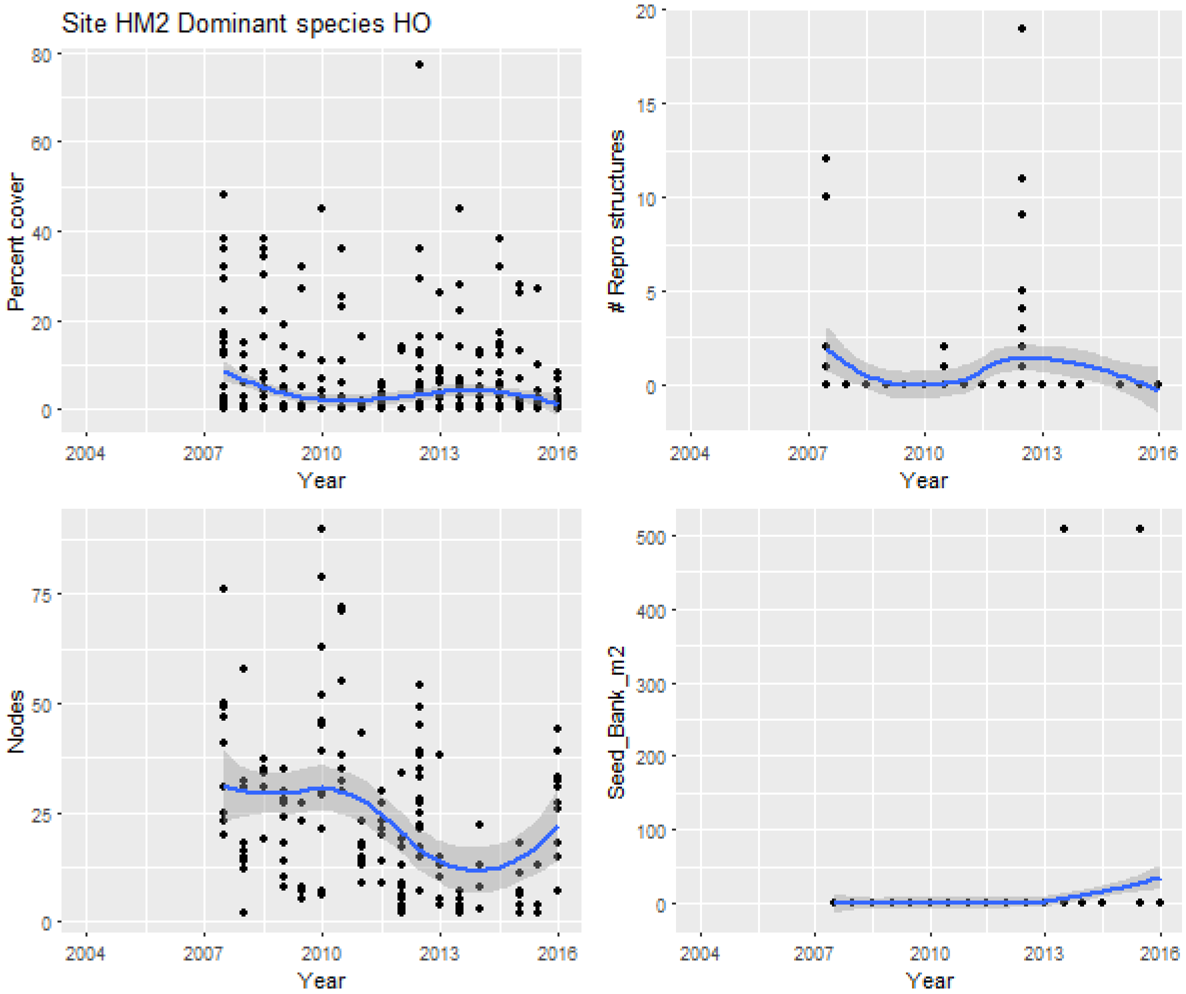

Figure A19: Summary of reproductive effort data for Hamilton Island 2, dominant species Halophila ovlais.

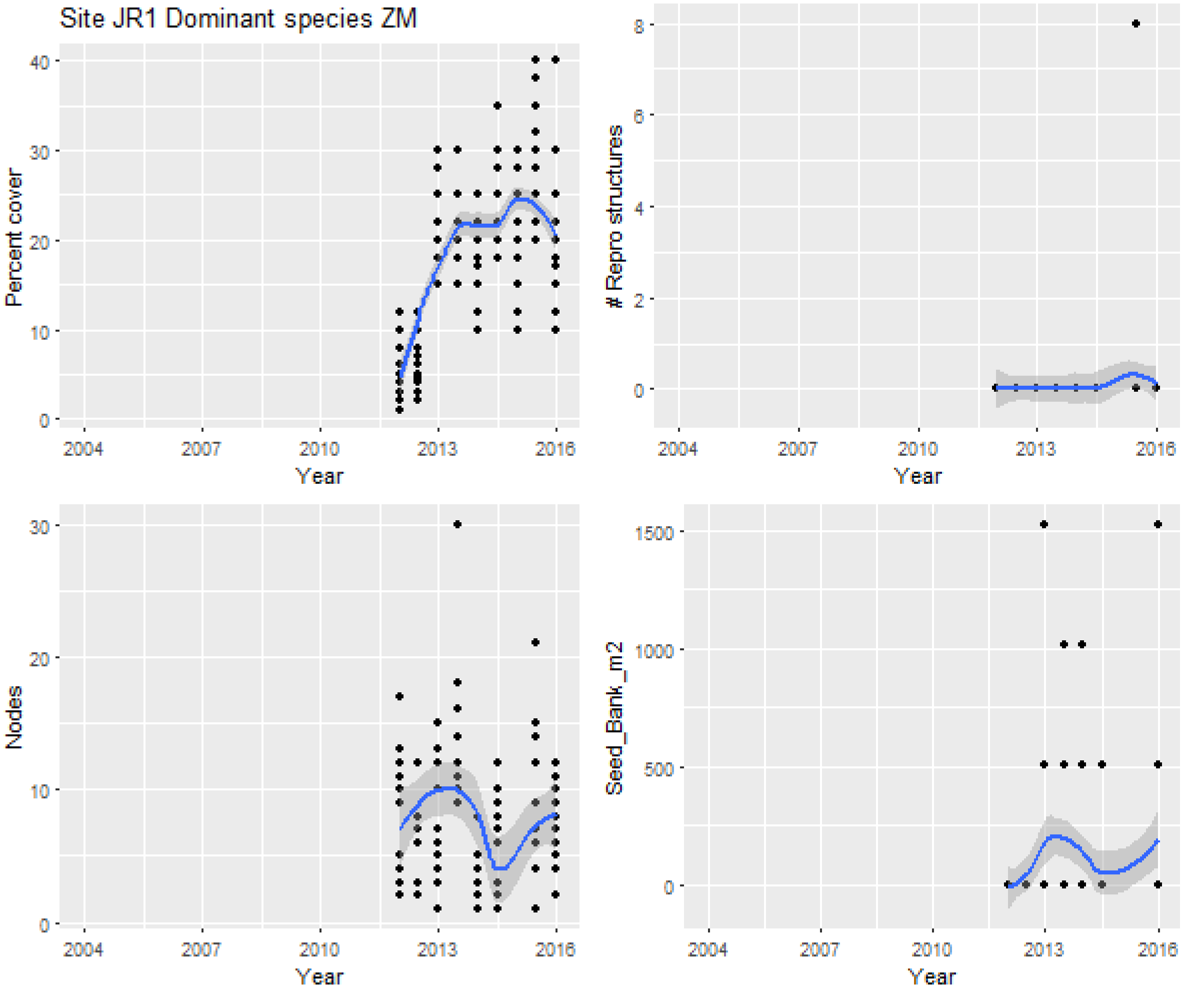

Figure A20: Summary of reproductive effort data for Jerona 1, dominant species Zostera.

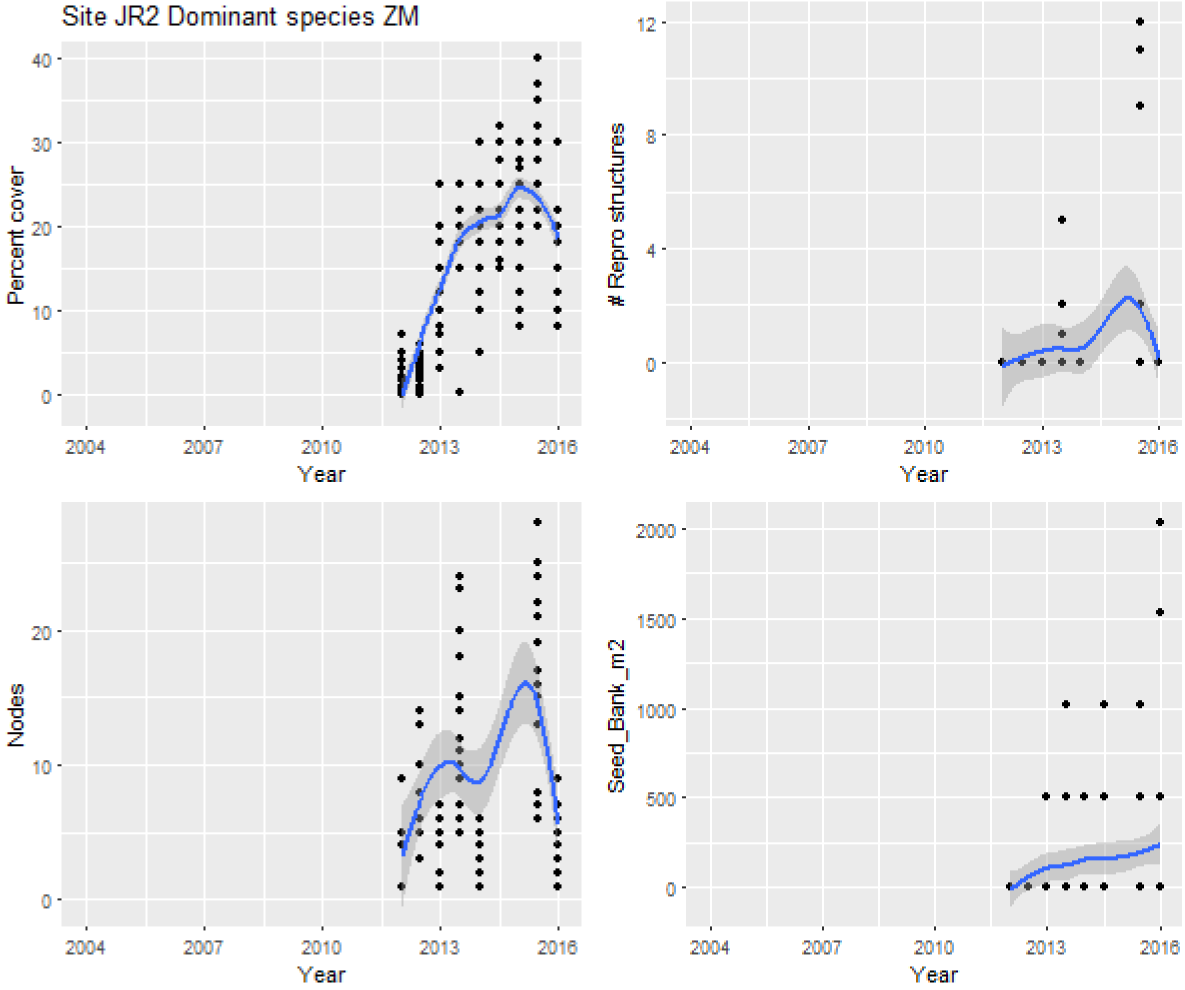

Figure A21: Summary of reproductive effort data for Jerona 2, dominant species Zostera.

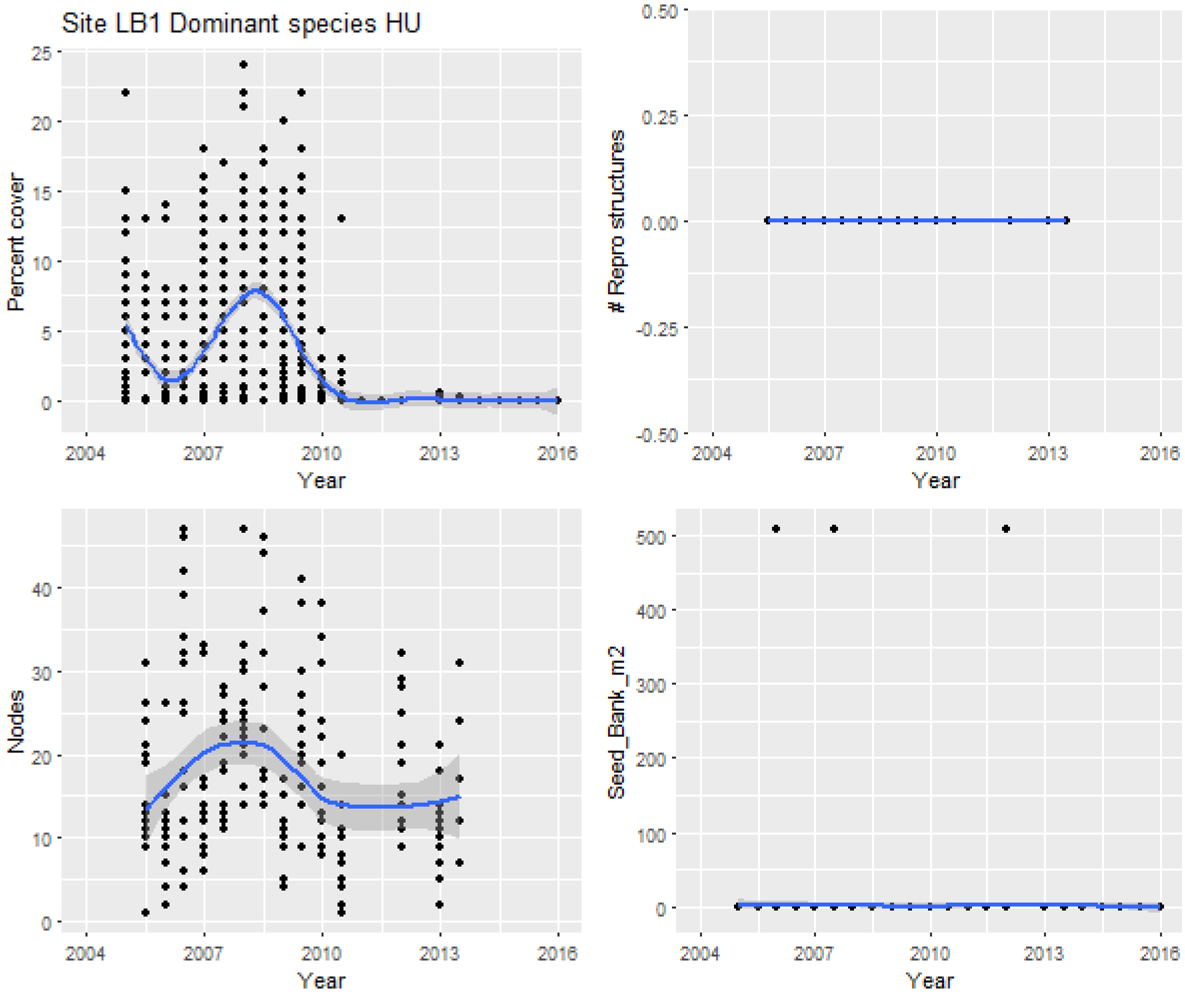

Figure A22: Summary of reproductive effort data for Lugger Bay 1, dominant species Halodule.

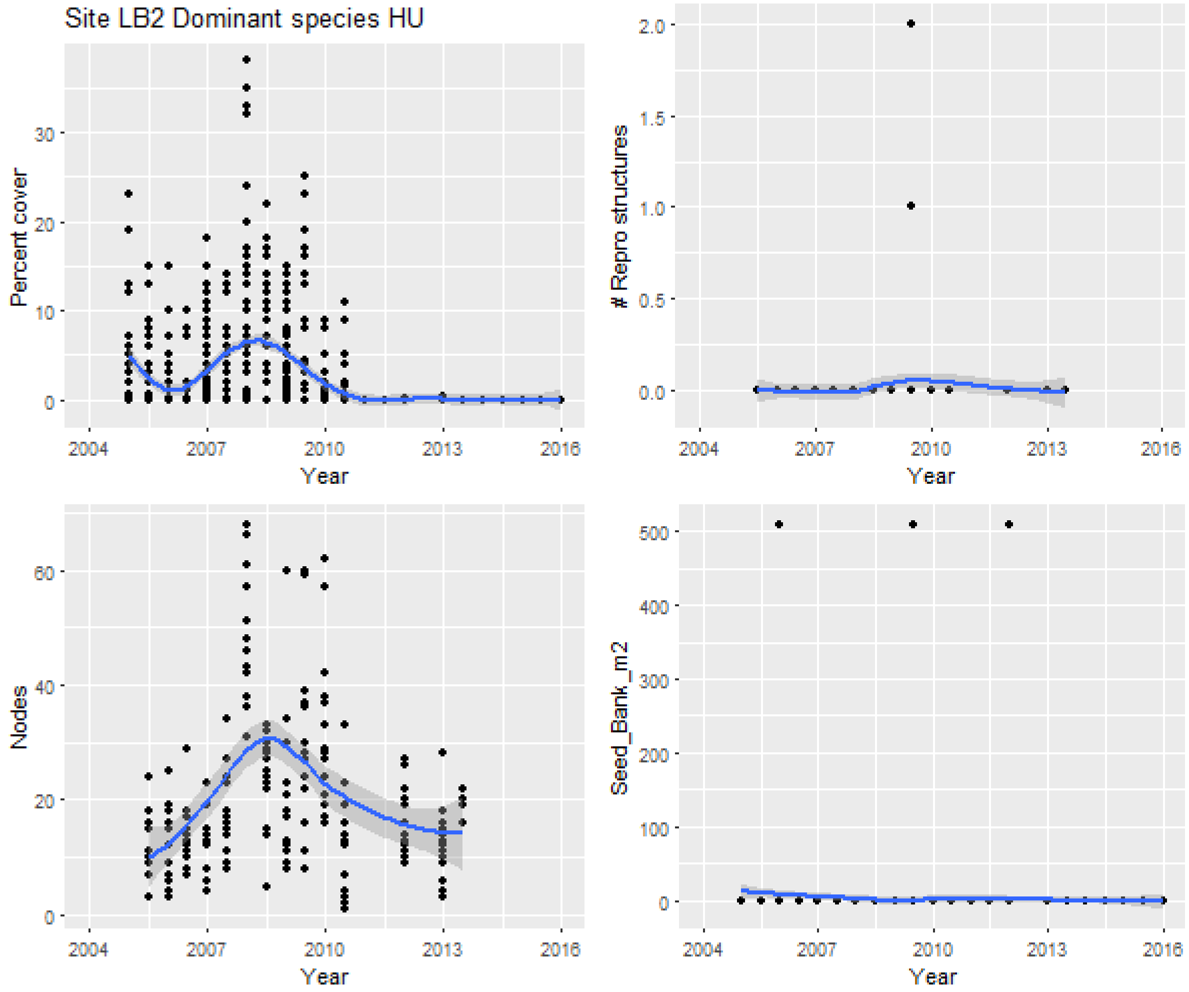

Figure A23: Summary of reproductive effort data for Lugger Bay 2, dominant species Halodule.

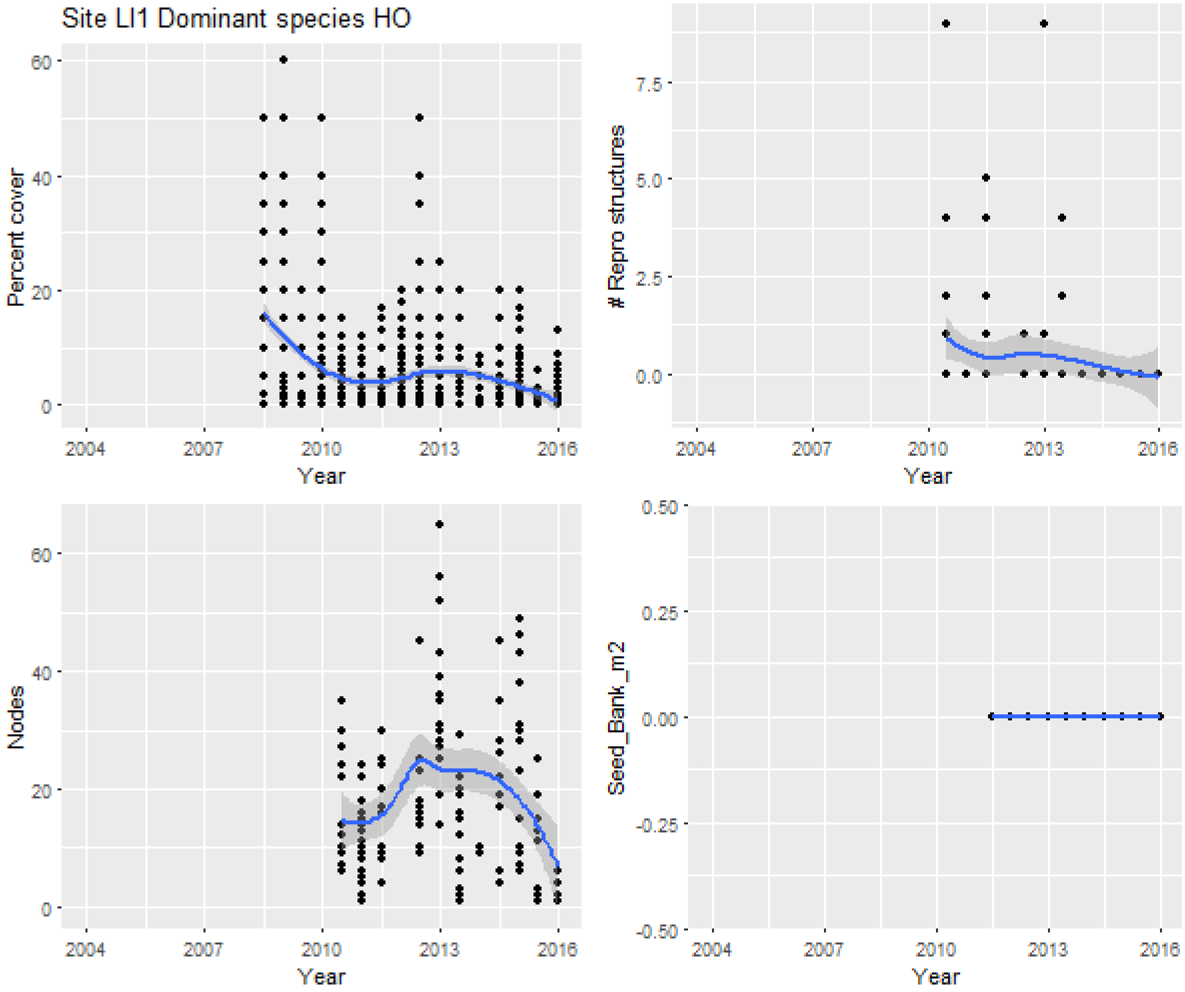

Figure A24: *Summary of reproductive effort data for Low Isles 1, dominant species Halophila ovalis.*

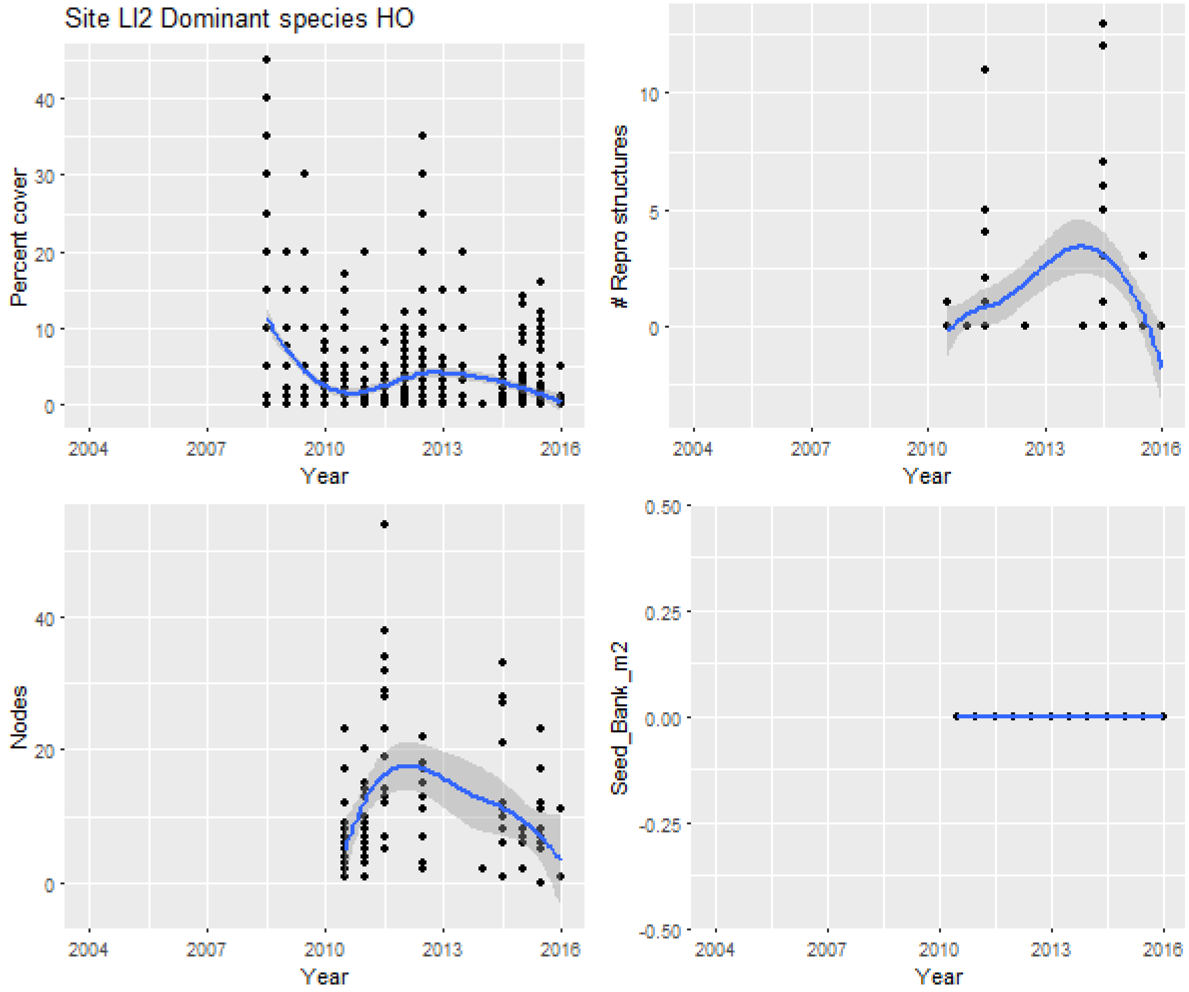

Figure A25: Summary of reproductive effort data for Low Isles 2, dominant species Halophila ovalis.

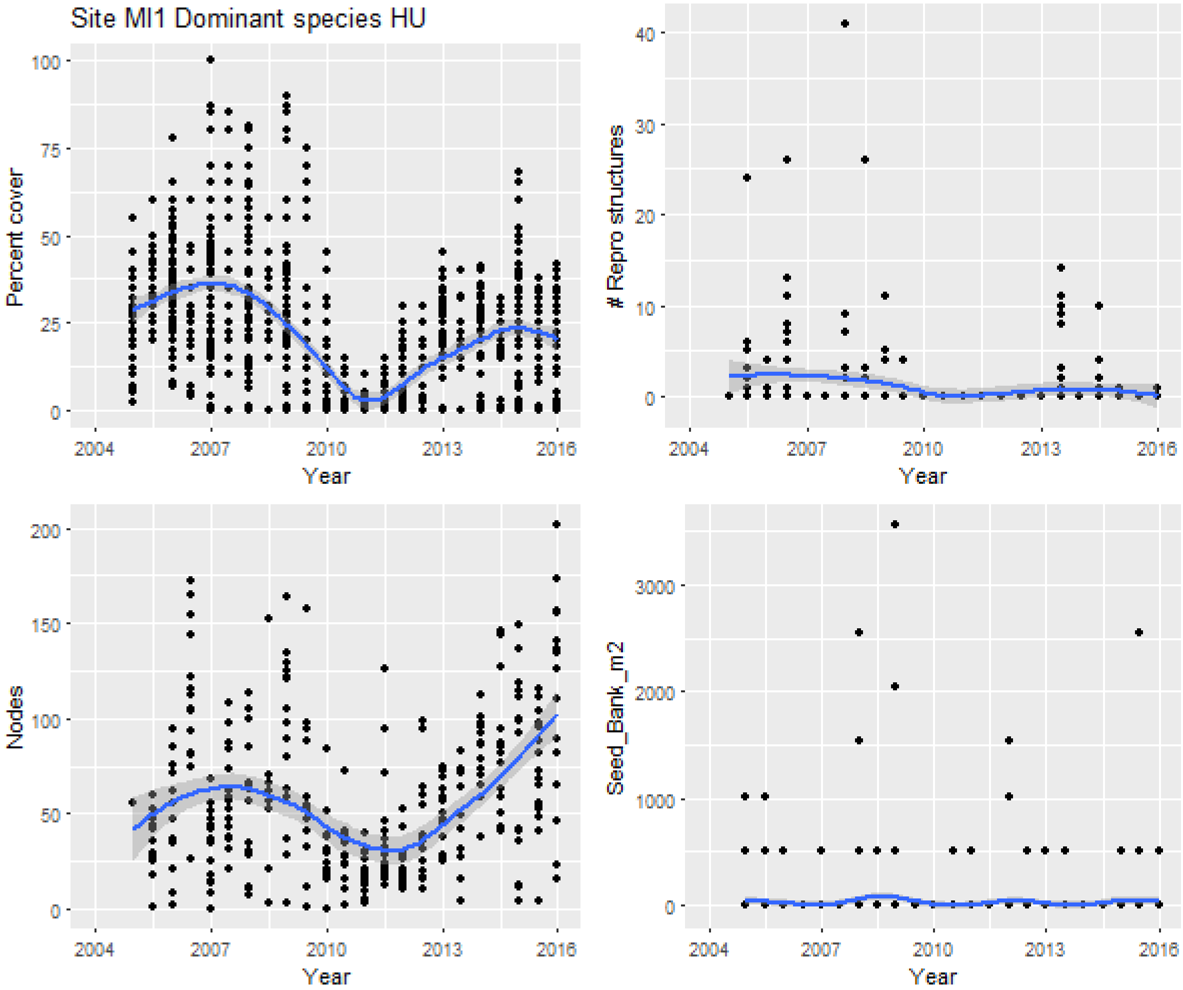

Figure A26: Summary of reproductive effort data for Picnic Bay 1, dominant species Halodule.

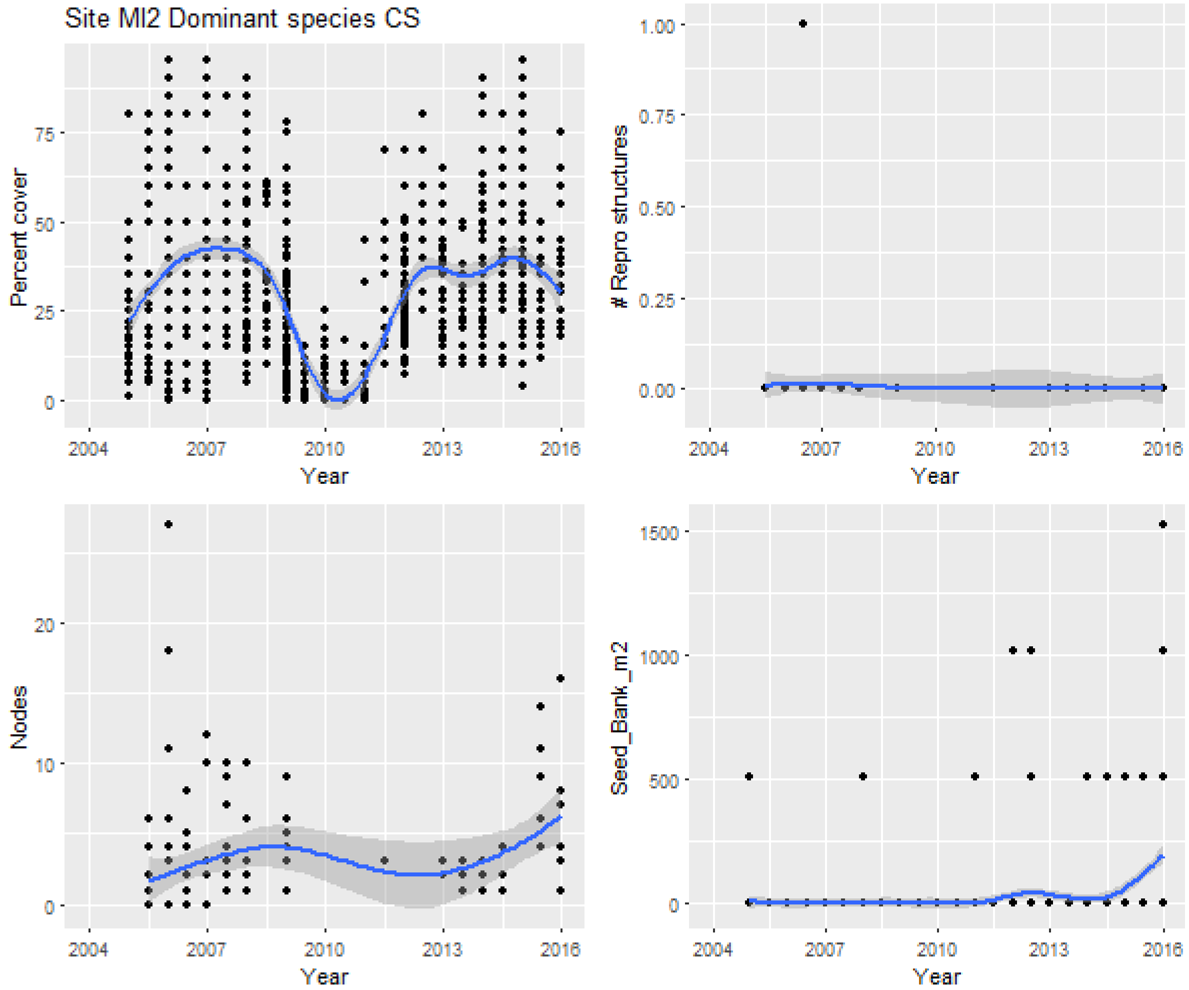

Figure A27: Summary of reproductive effort data for Cockle Bay, dominant species Cymodocea serrulata.

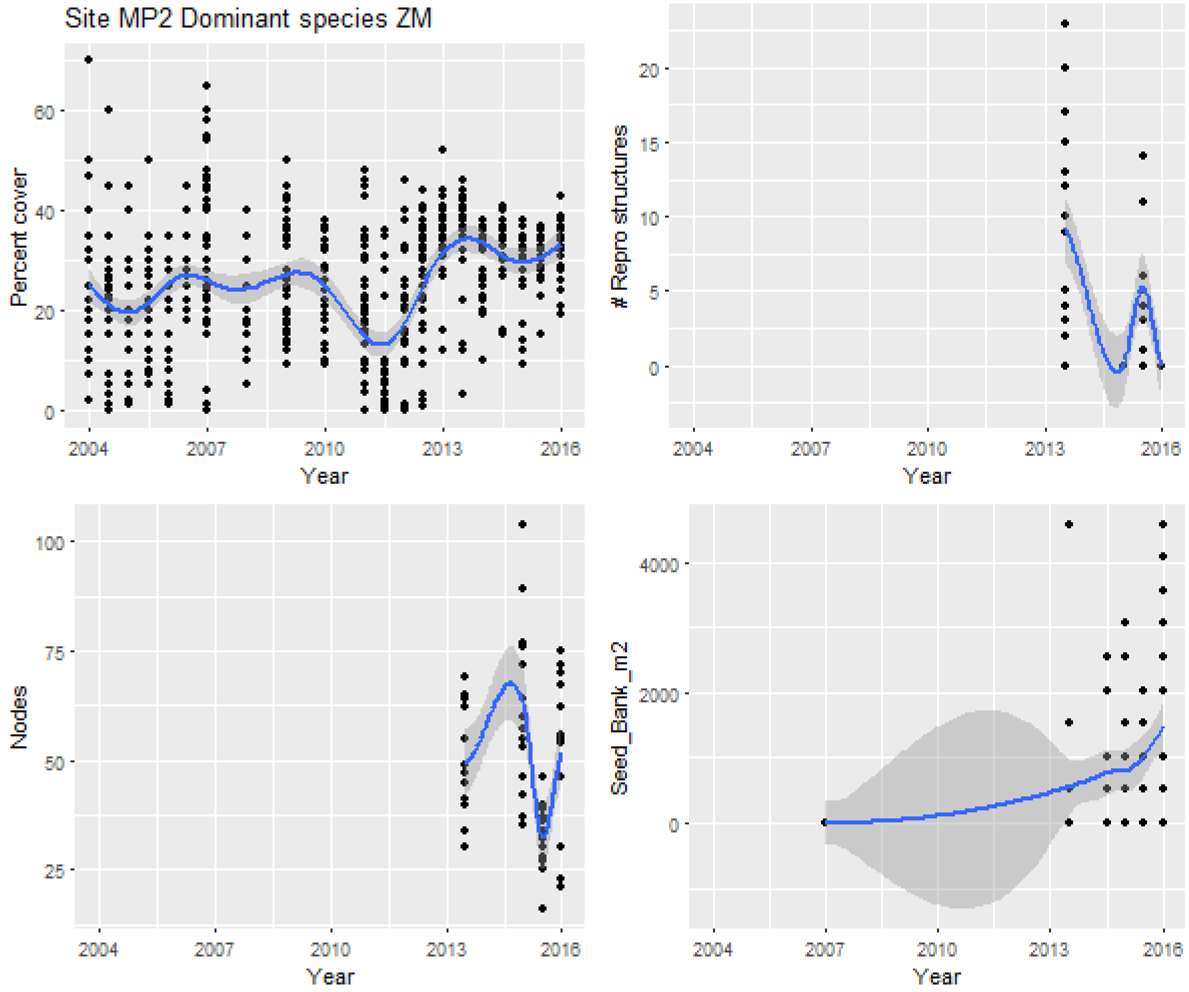

Figure A28: Summary of reproductive effort data for Midge Point 2, dominant species Zostera.

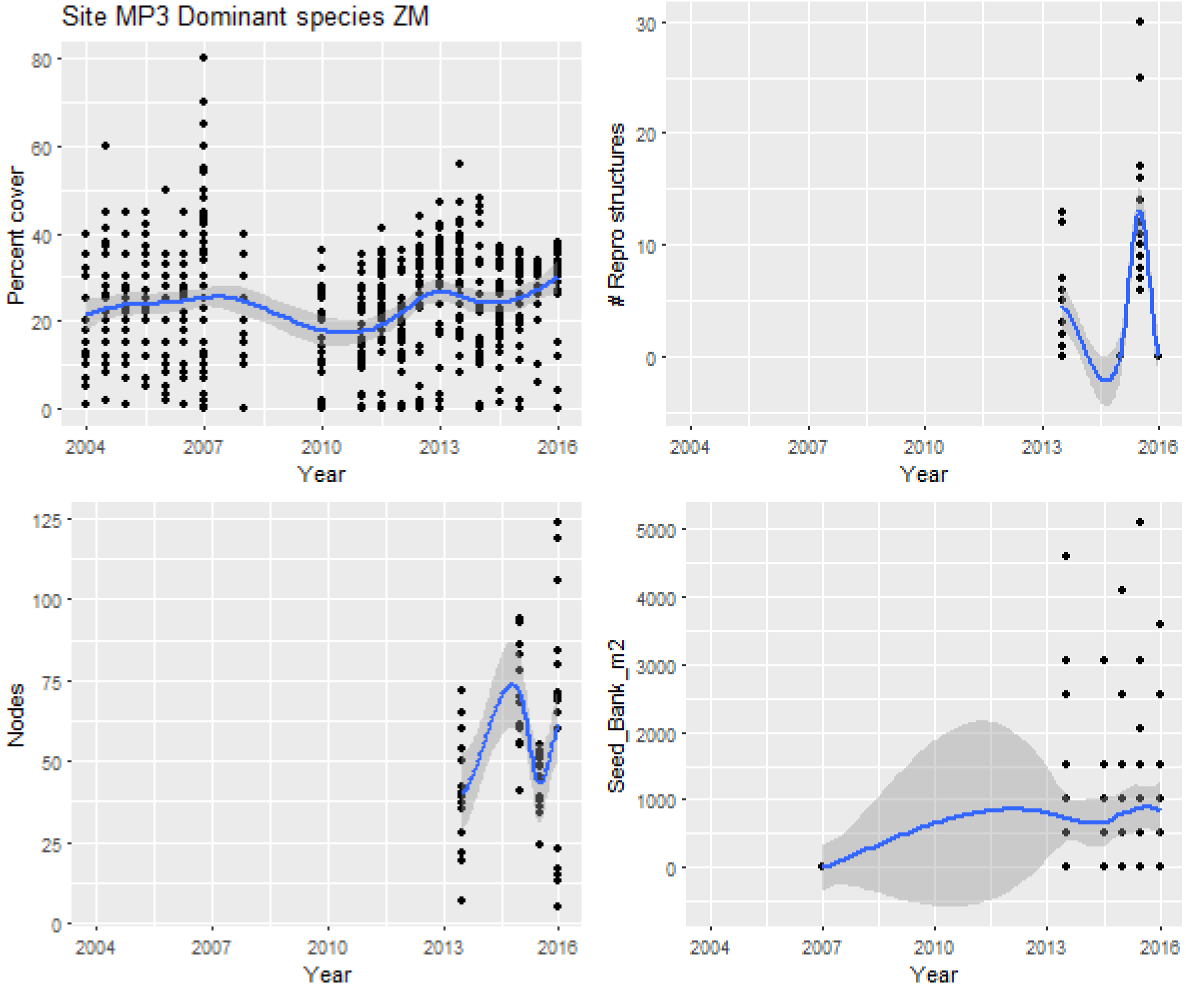

Figure A29: Summary of reproductive effort data for Midge Point 2, dominant species Zostera.

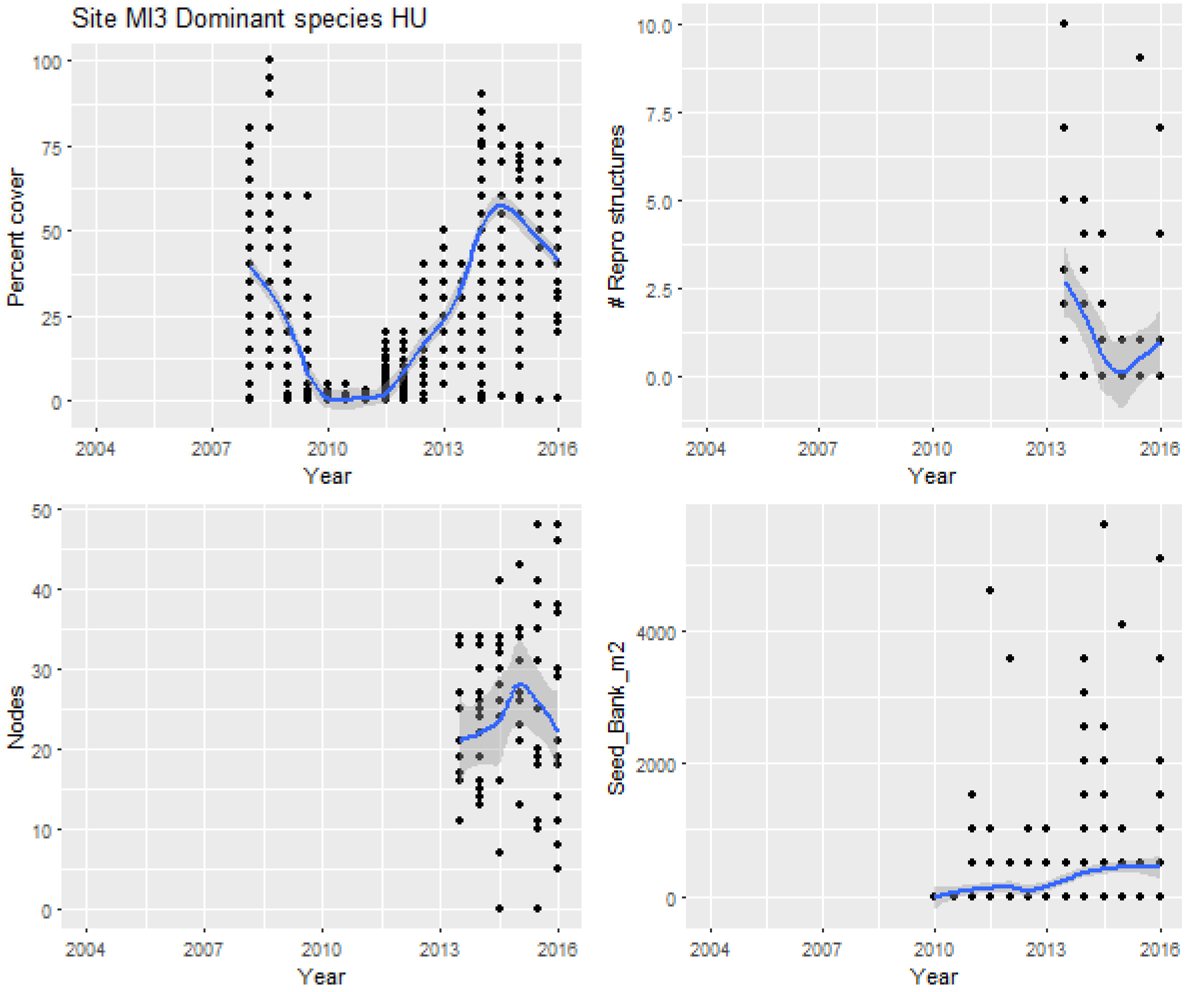

Figure A30: Summary of reproductive effort data for Picnic Bay 2, dominant species Halodule.

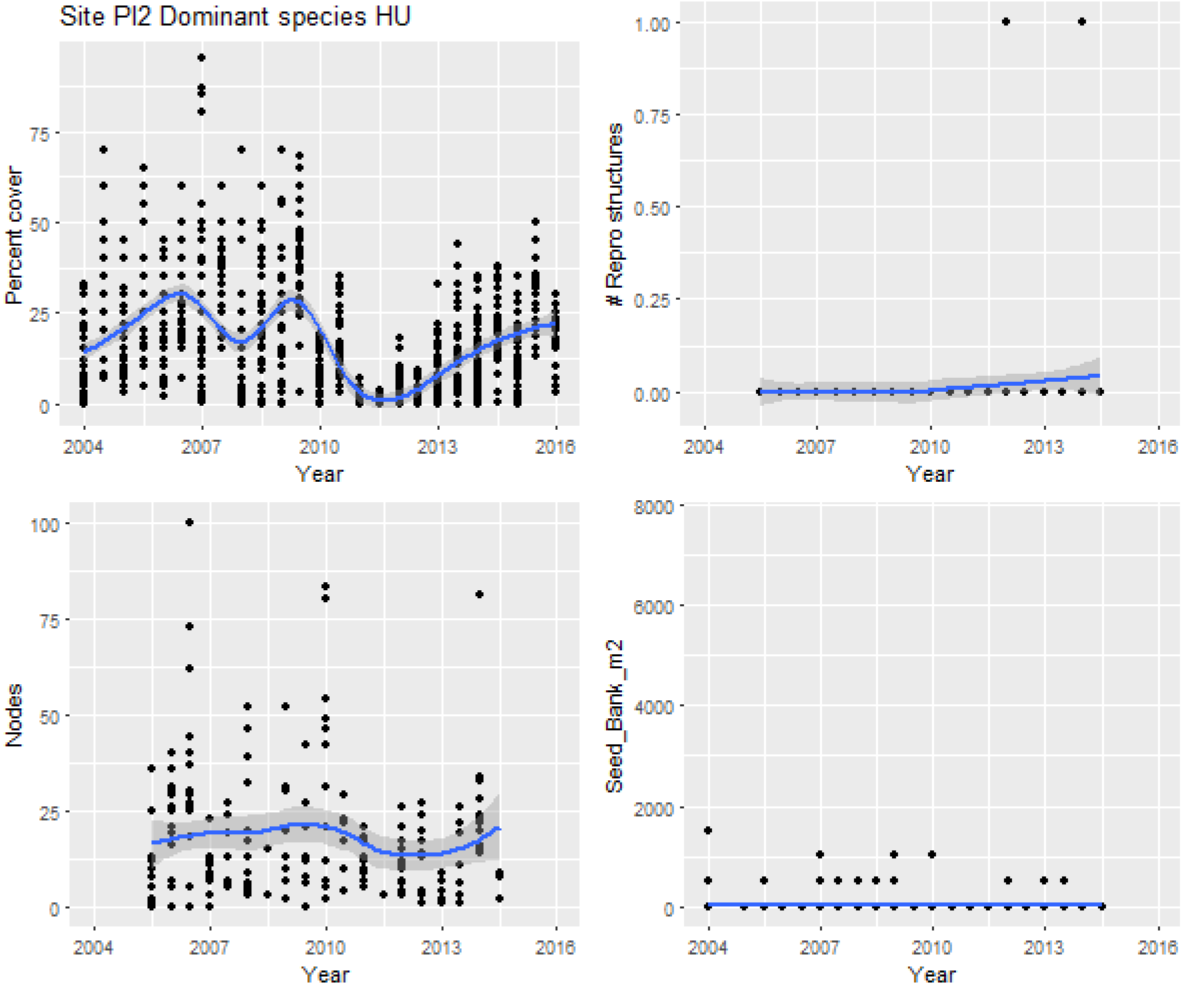

Figure A31: Summary of reproductive effort data for Pioneer Bay 2, dominant species Halodule.

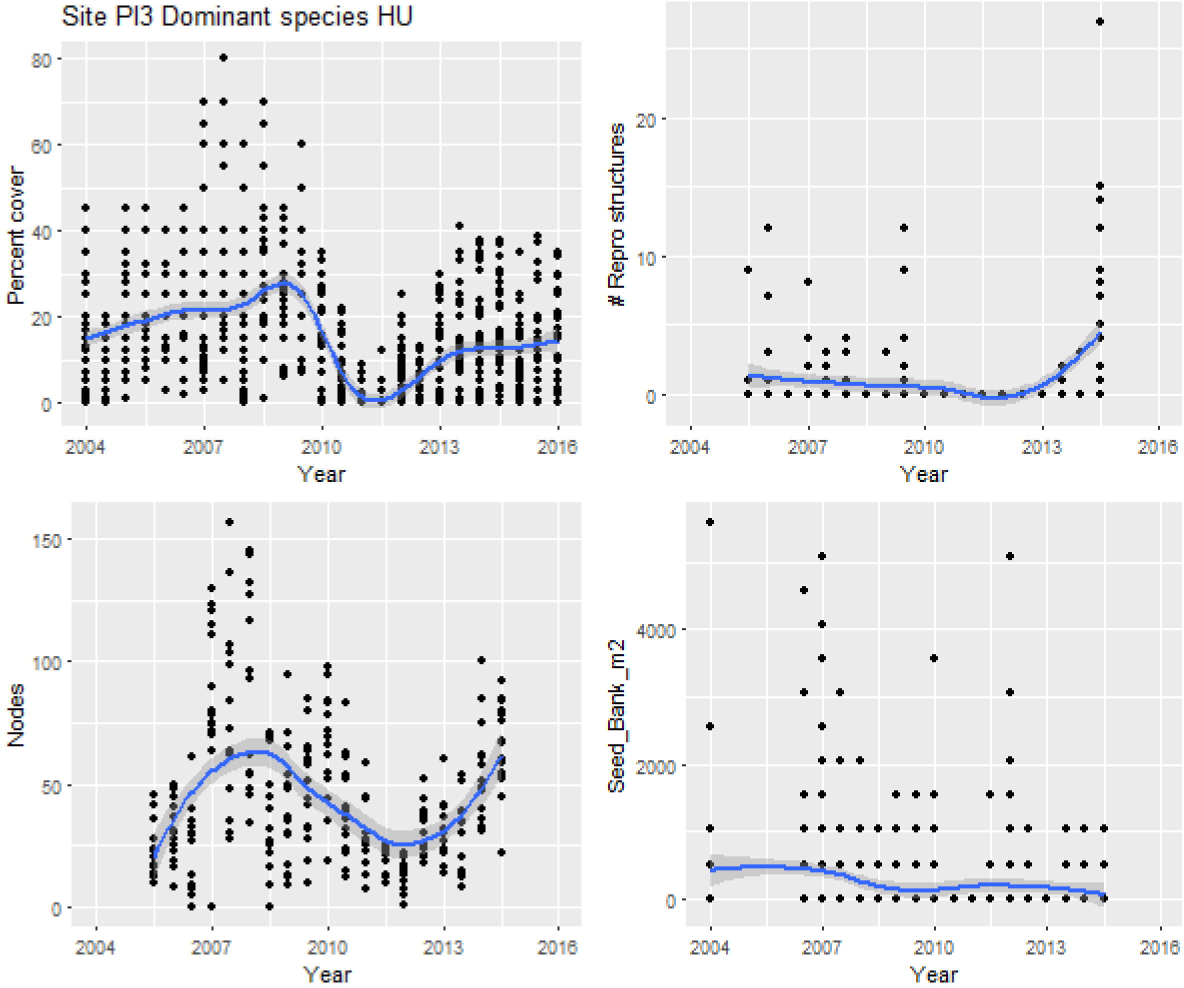

Figure A32: Summary of reproductive effort data for Pioneer Bay 2, dominant species Halodule.

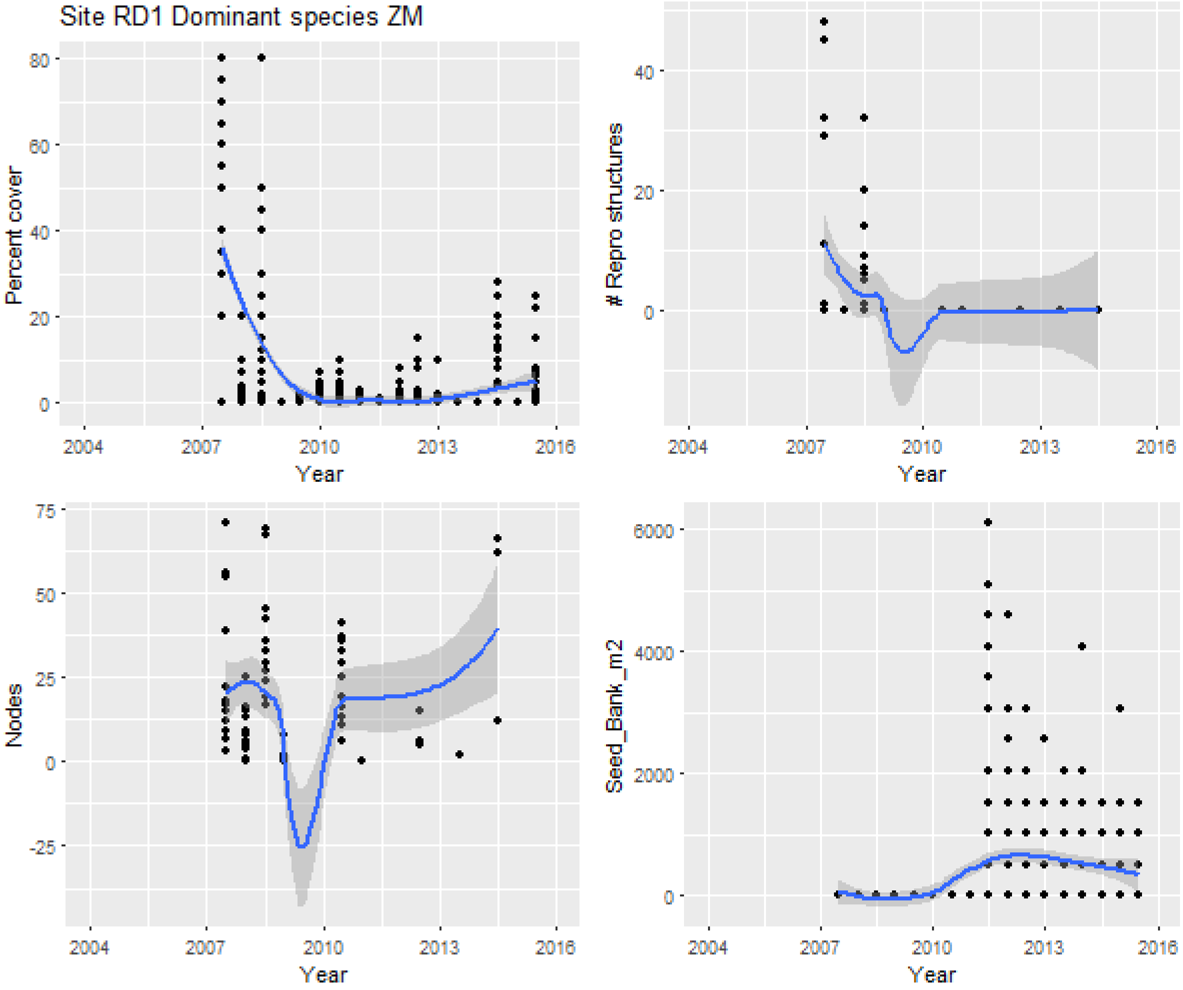

Figure A33: Summary of reproductive effort data for Rods Bay 1, dominant species Zostera.

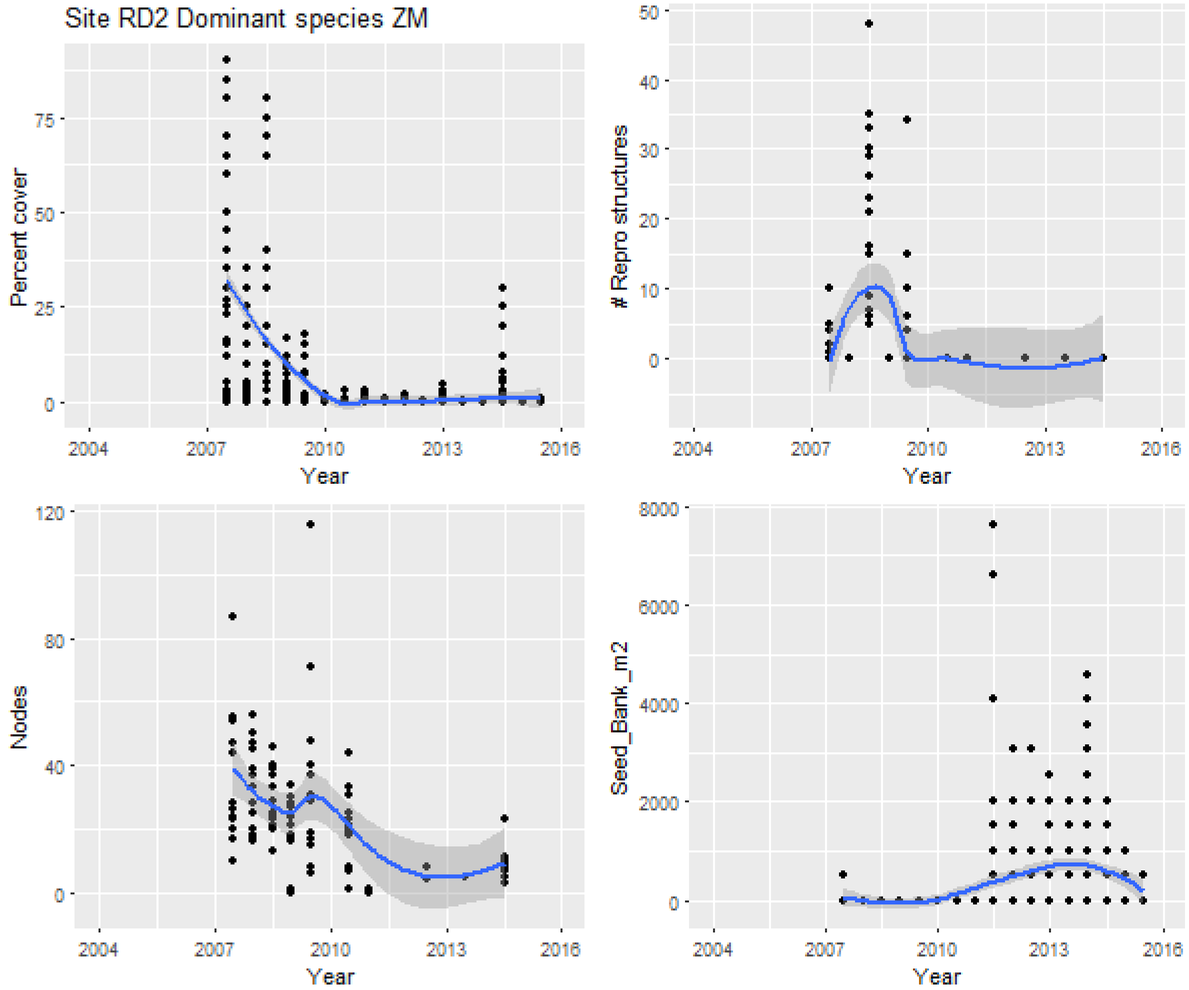

Figure A34: Summary of reproductive effort data for Rodds Bay 2, dominant species Zostera.

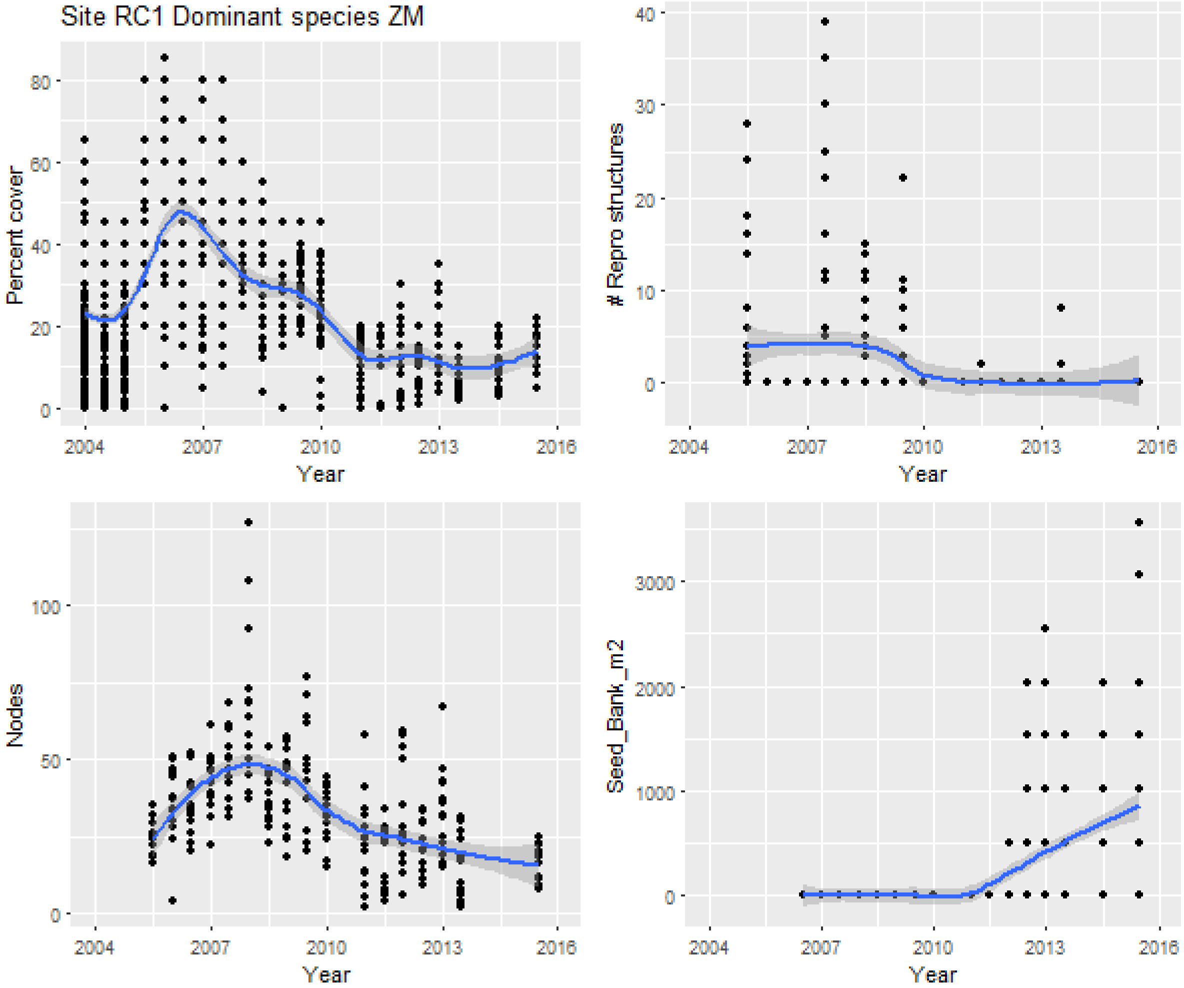

Figure A35: Summary of reproductive effort data for Ross Creek 1, dominant species Zostera.

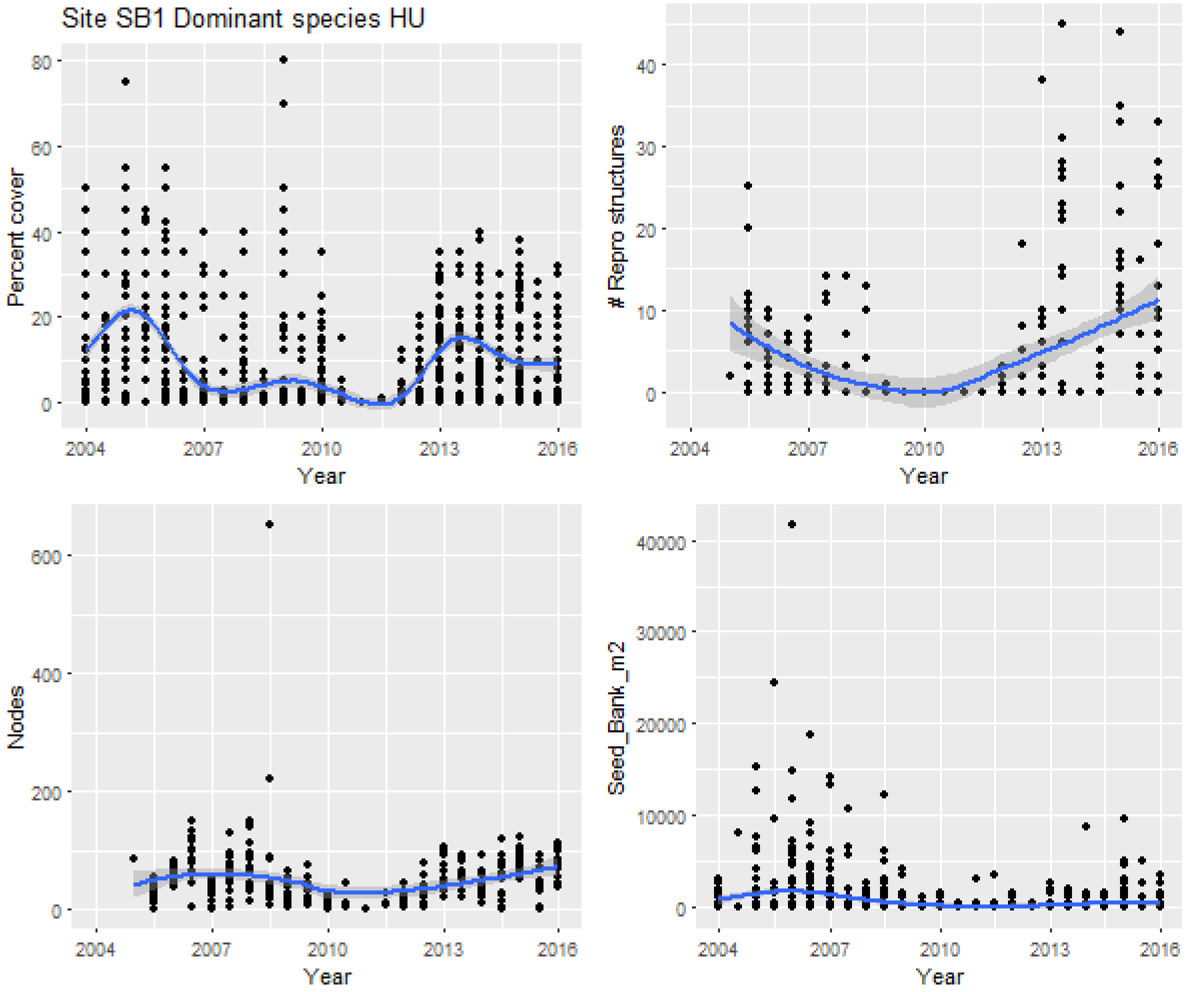

Figure A36: Summary of reproductive effort data for Shelley Beach 1, dominant species Halodule.

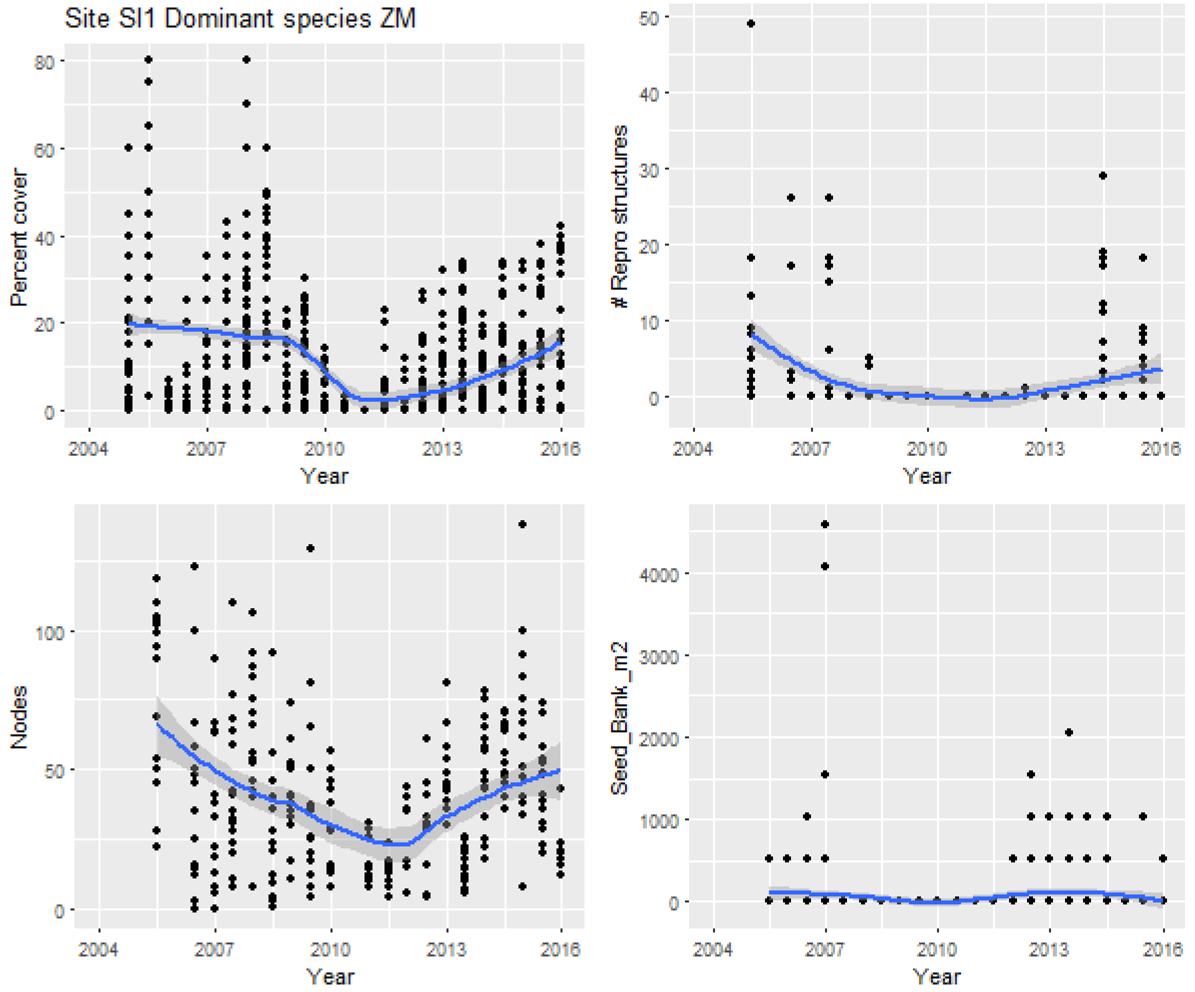

Figure A37: Summary of reproductive effort data for Sarina Inlet 1, dominant species Zostera.

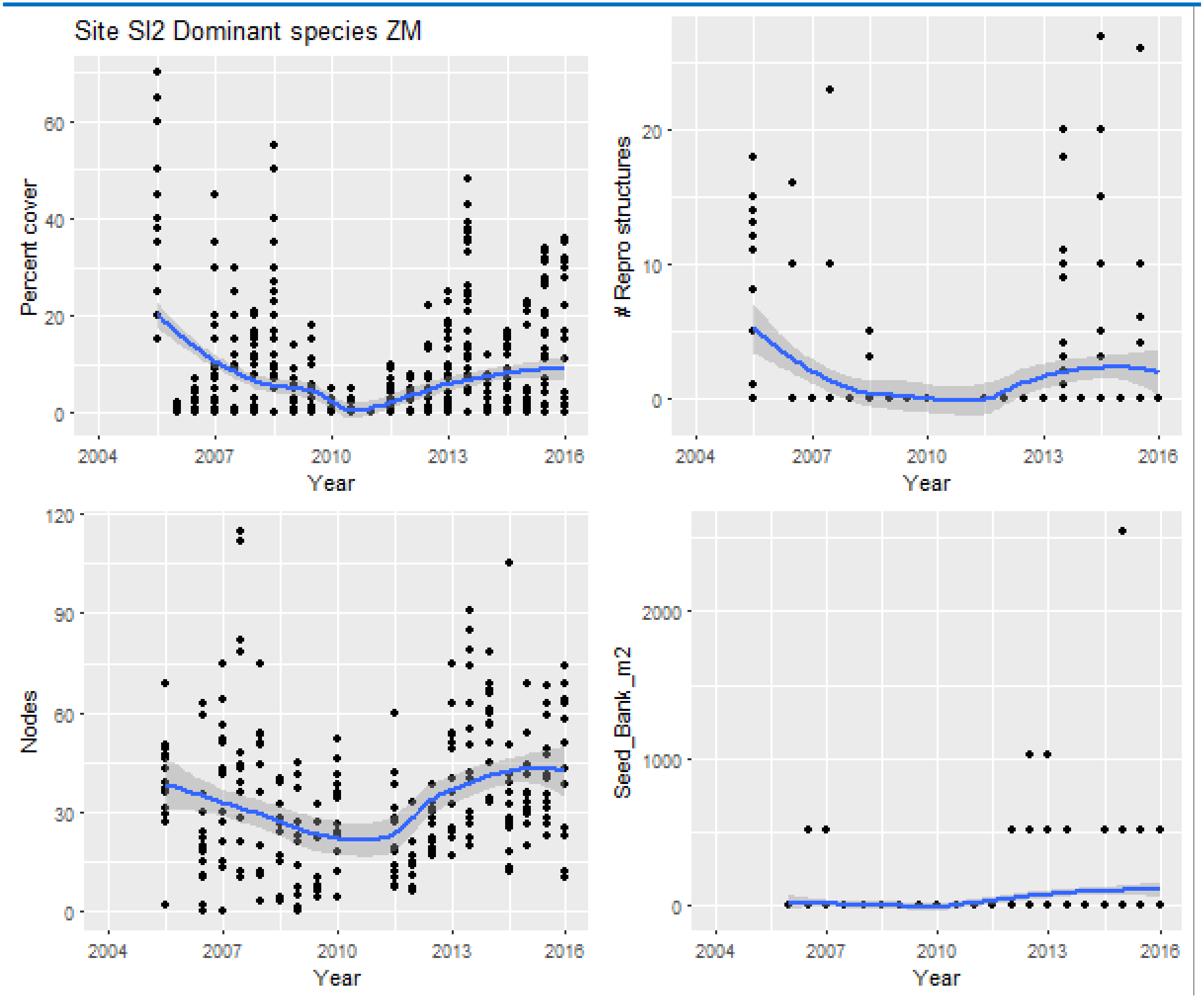

Figure A38: Summary of reproductive effort data for Sarina Inlet 2, dominant species Zostera.

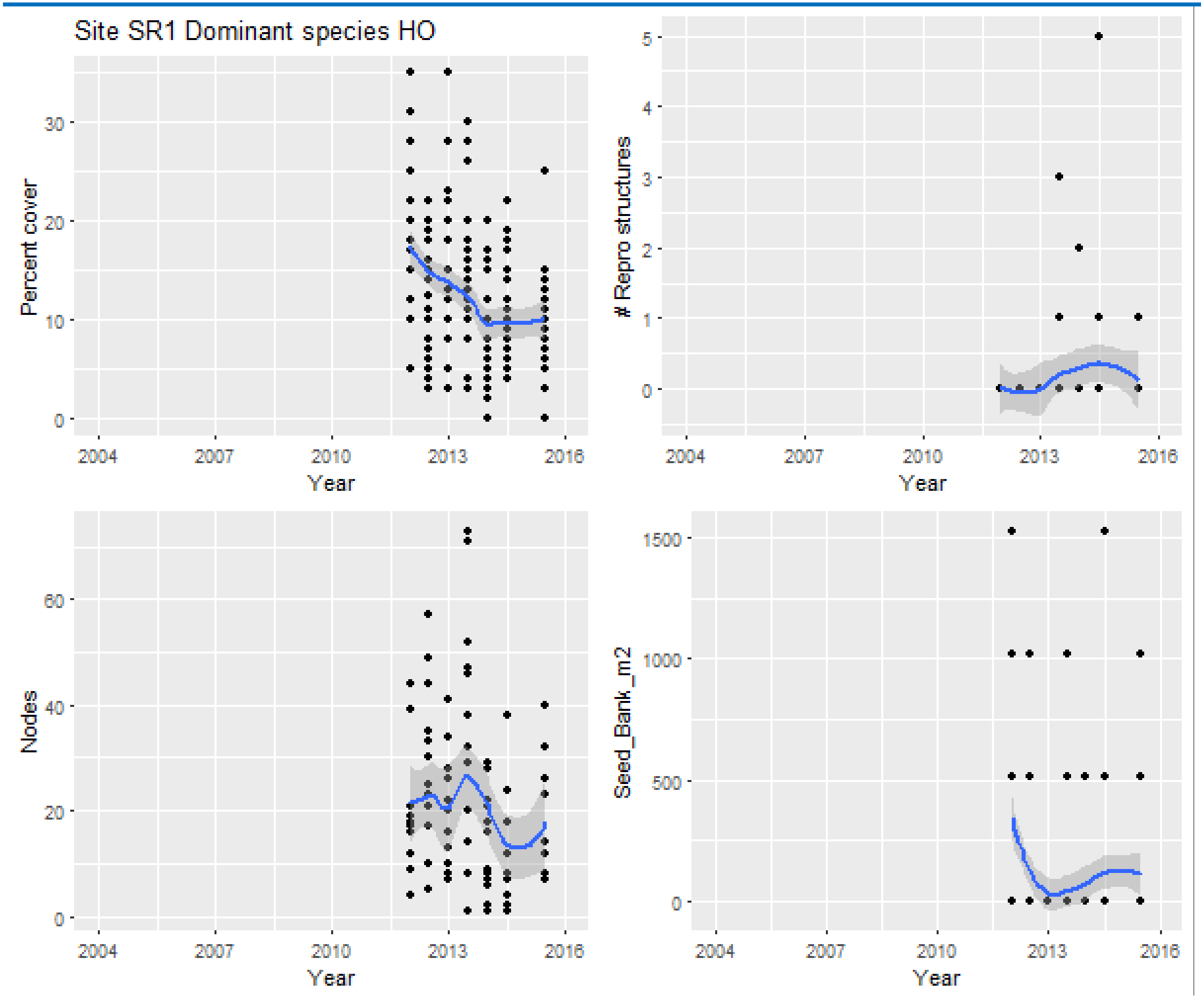

Figure A39: Summary of reproductive effort data for Shellburne Bay 1, dominant species Halophila ovalis.

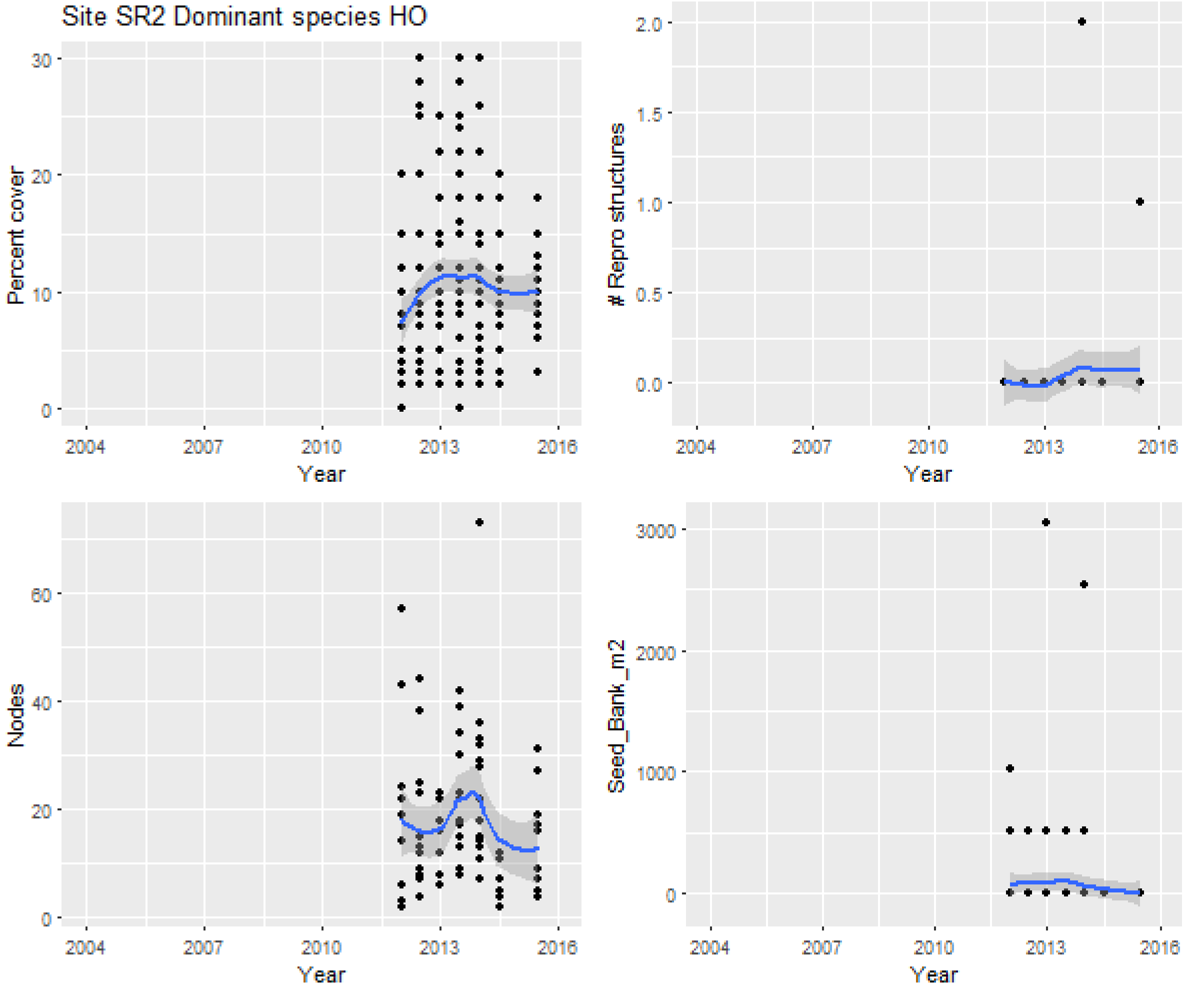

Figure A40: Summary of reproductive effort data for Shellburne Bay 2, dominant species Halophila ovalis.

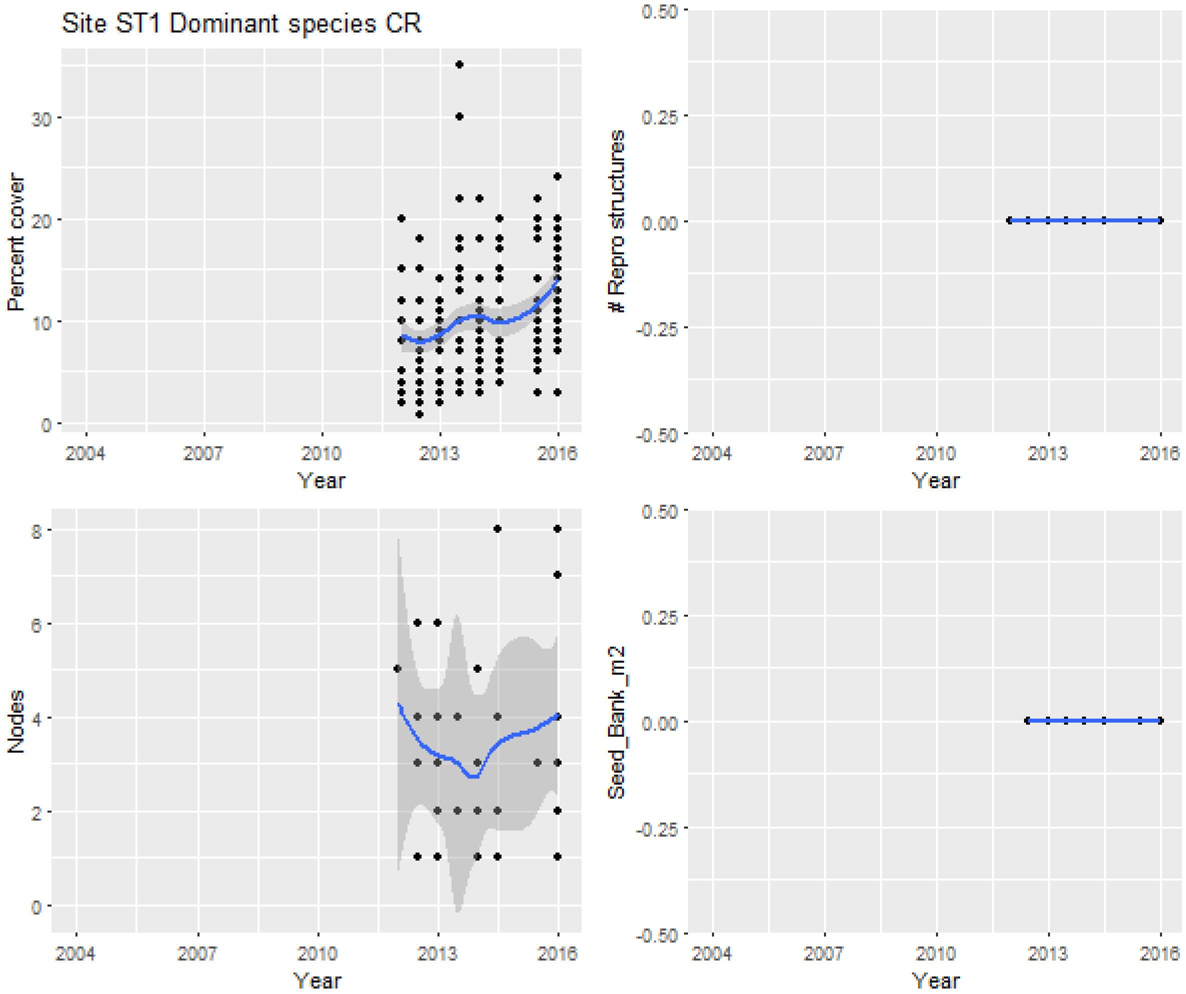

Figure A41: Summary of reproductive effort data for Stanley Island 1, dominant species Cymodocea rotundata.

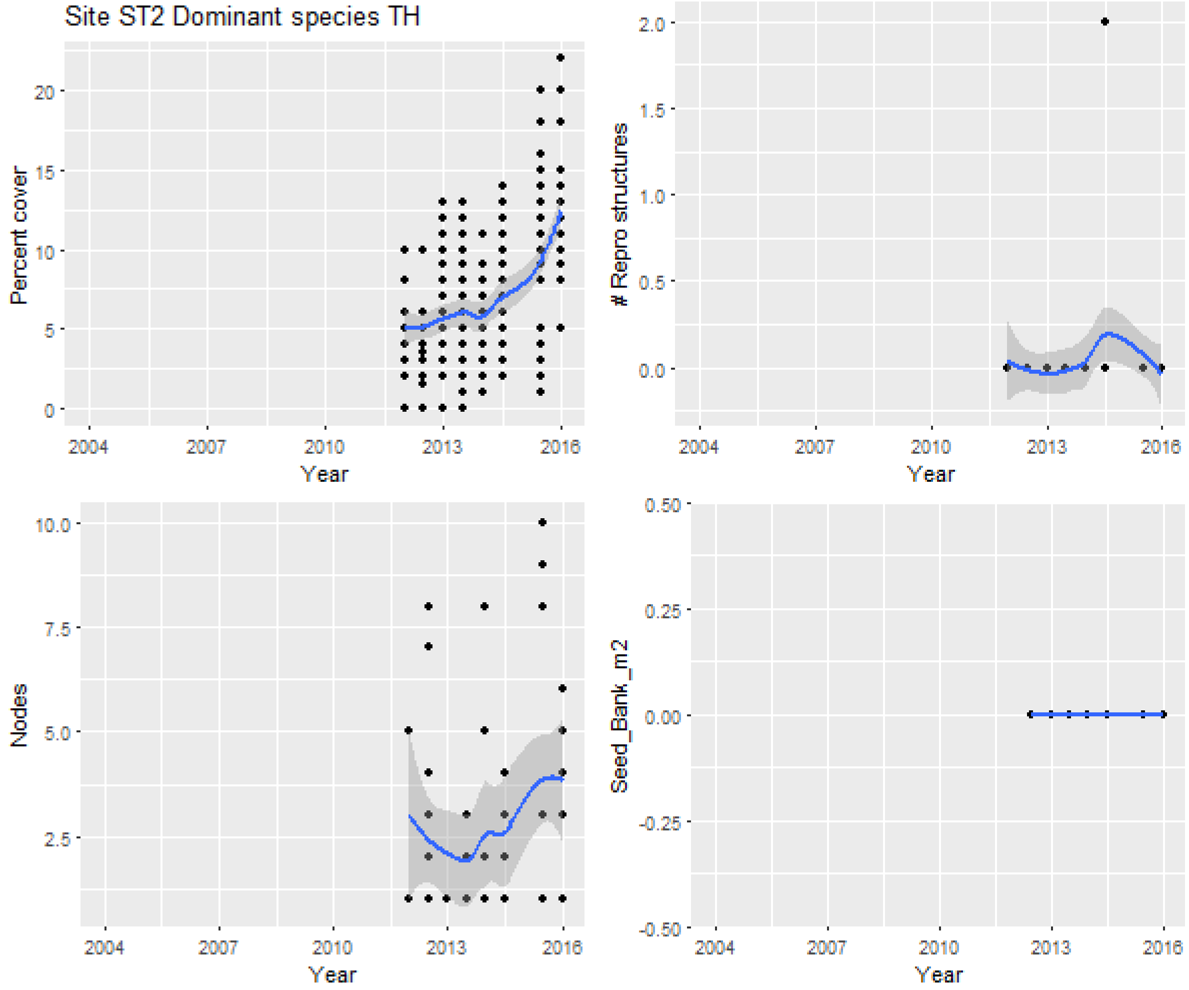

Figure A42: Summary of reproductive effort data for Stanley Island 1, dominant species Thalassia.

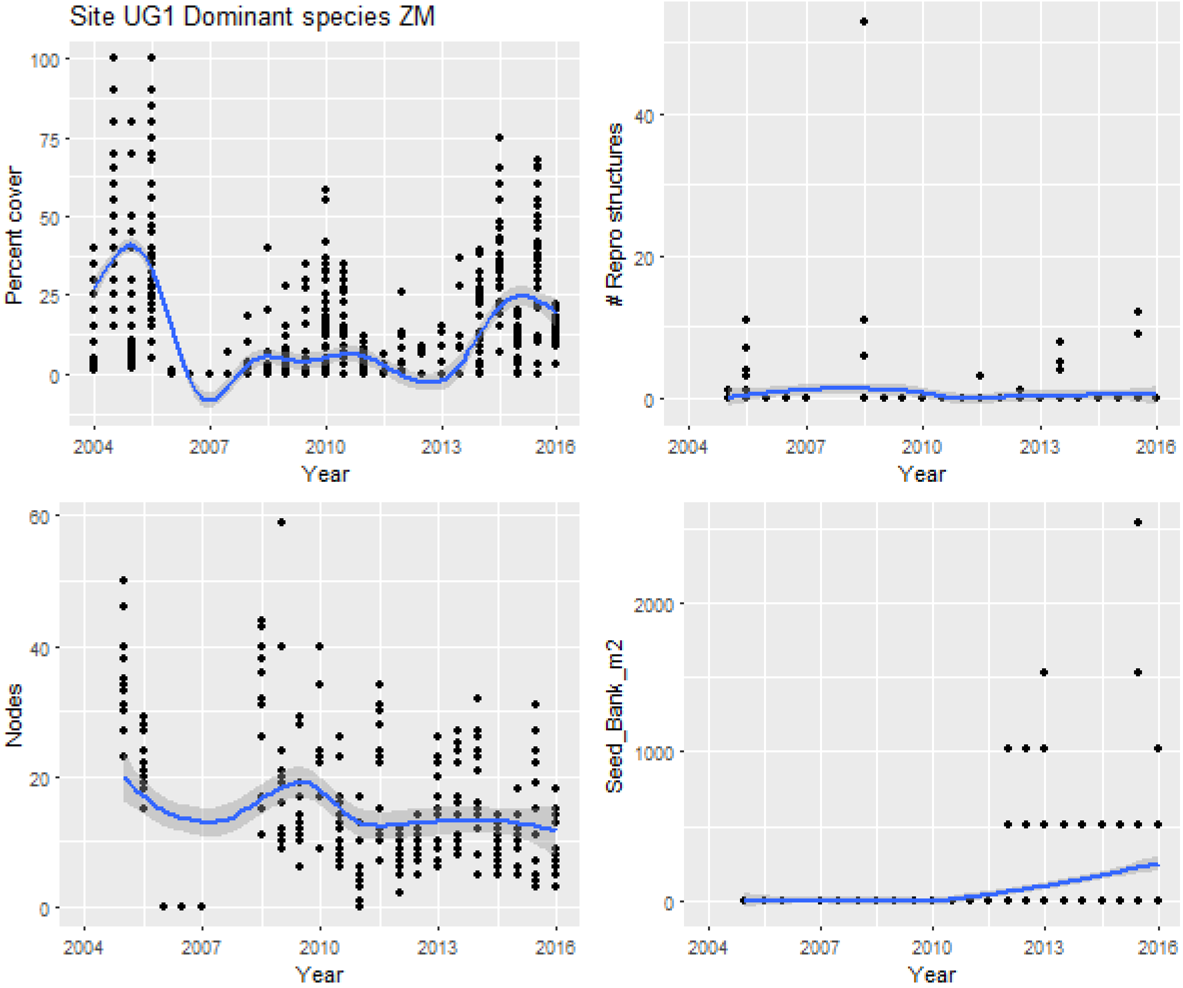

Figure A43: Summary of reproductive effort data for Urangan 1, dominant species Zostera.

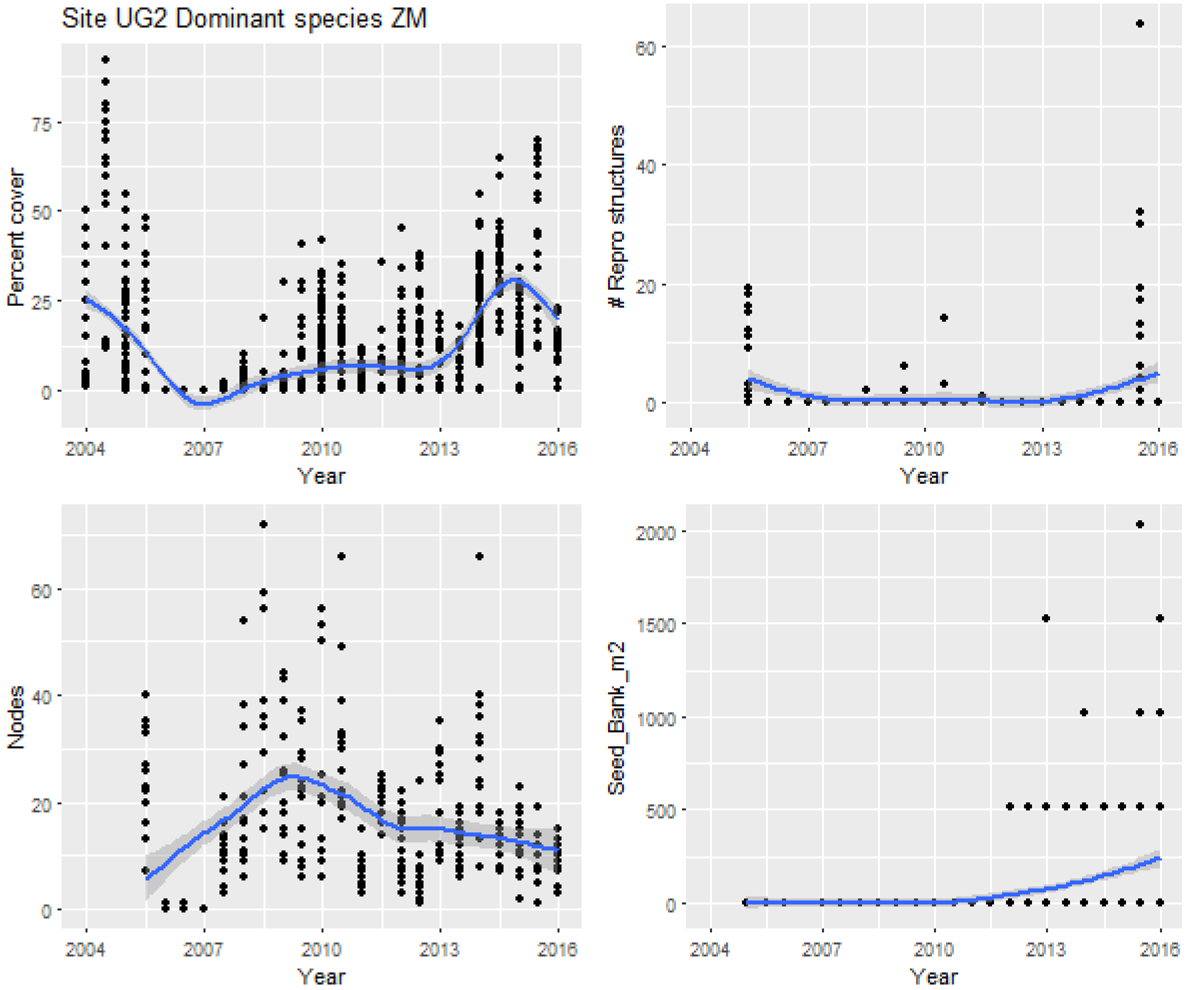

Figure A44: Summary of reproductive effort data for Urangan 2, dominant species Zostera.

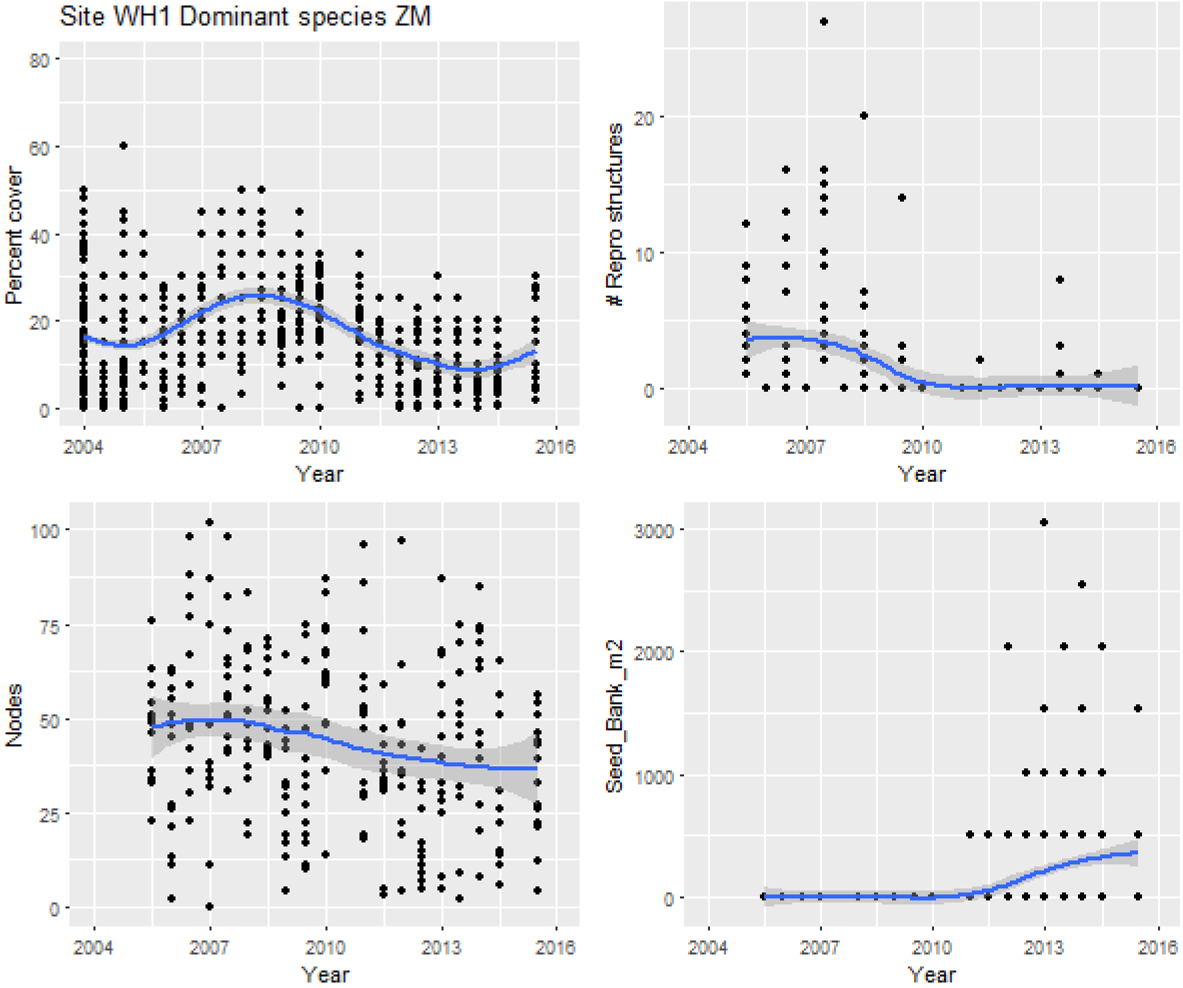

Figure A45: Summary of reproductive effort data for Whellens hut dominant species Zostera.

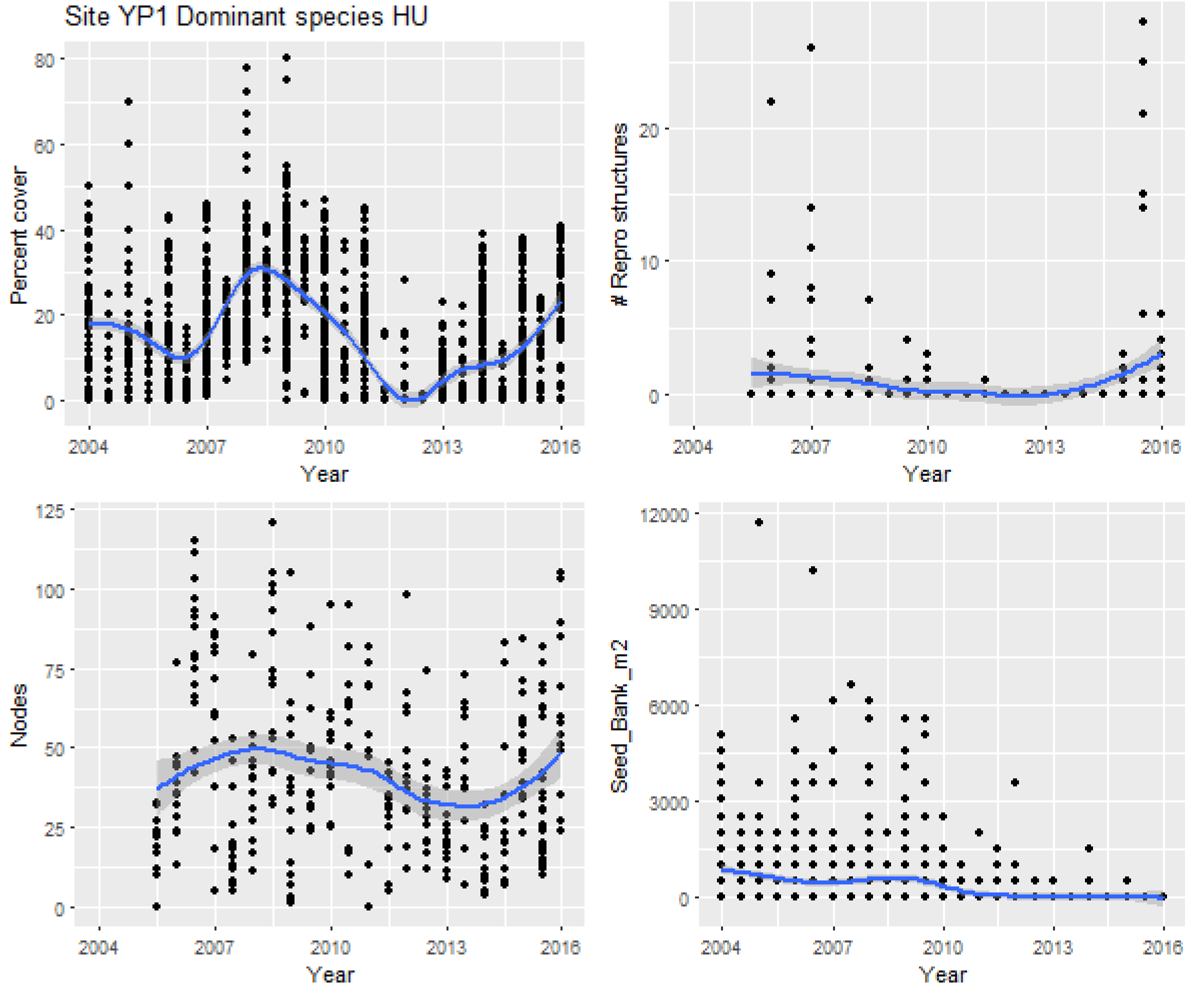

Figure A46: Summary of reproductive effort data for Yule Point 1, dominant species Halodule.

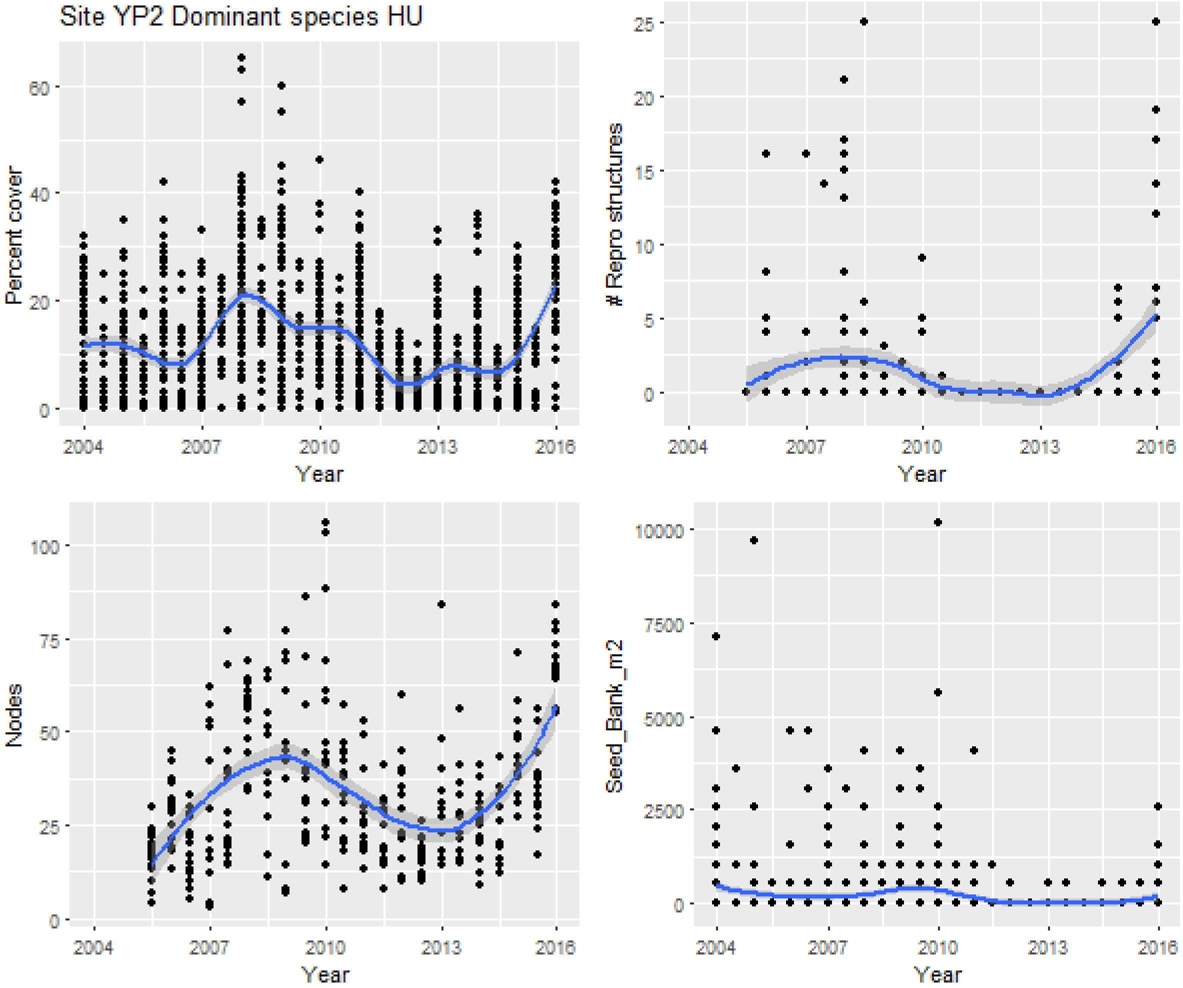

Figure A47: Summary of reproductive effort data for Yule Point 2, dominant species Halodule.

Appendix B. Model outputs

Observed mean seagrass vs zero-inflated beta regression and random forest model predictions

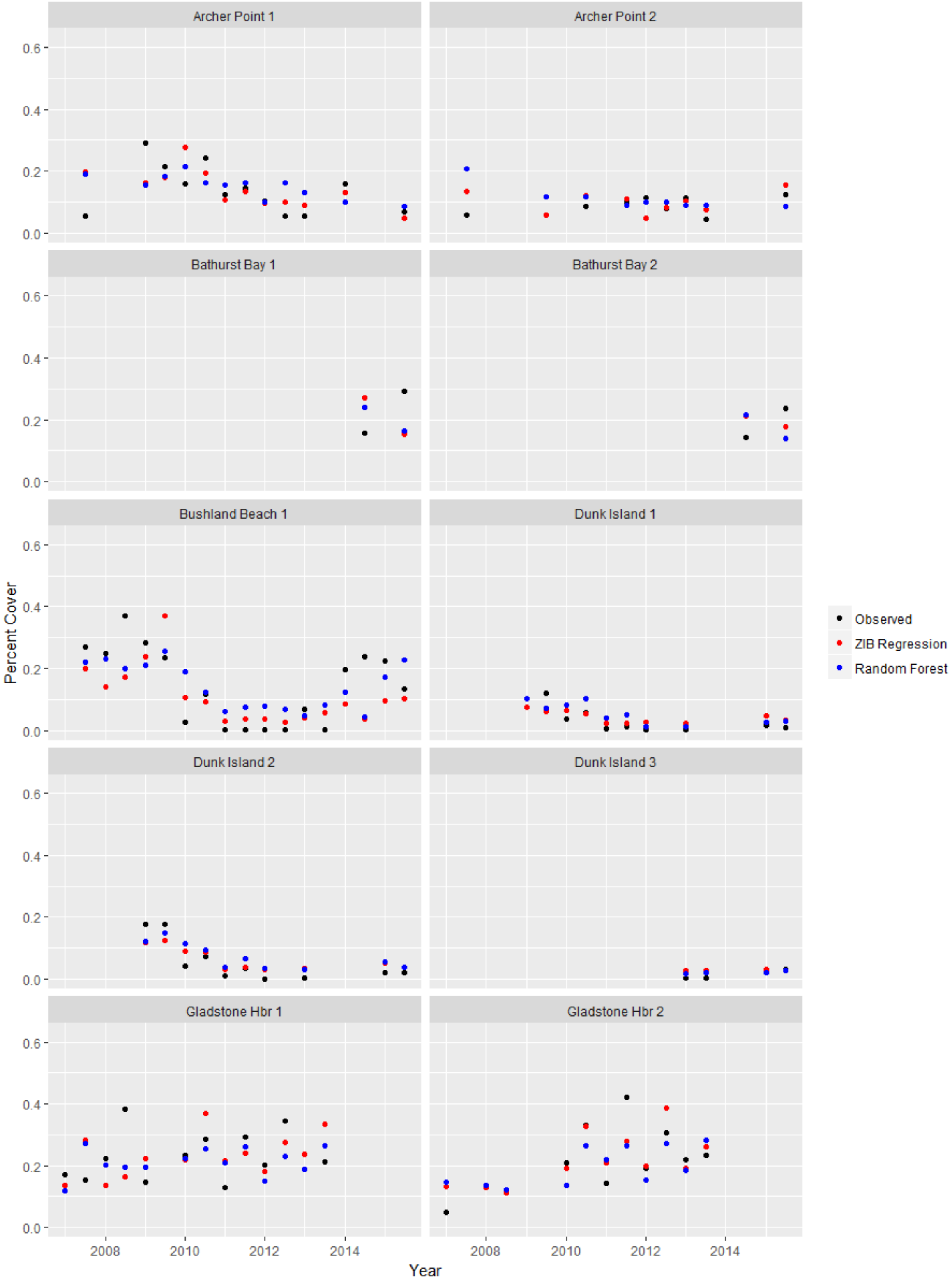

Figure B1: Observed (black) and predicted percent cover of seagrass for the zero inflated beta regression (red) and random forest (blue) for individual sites using model M4.

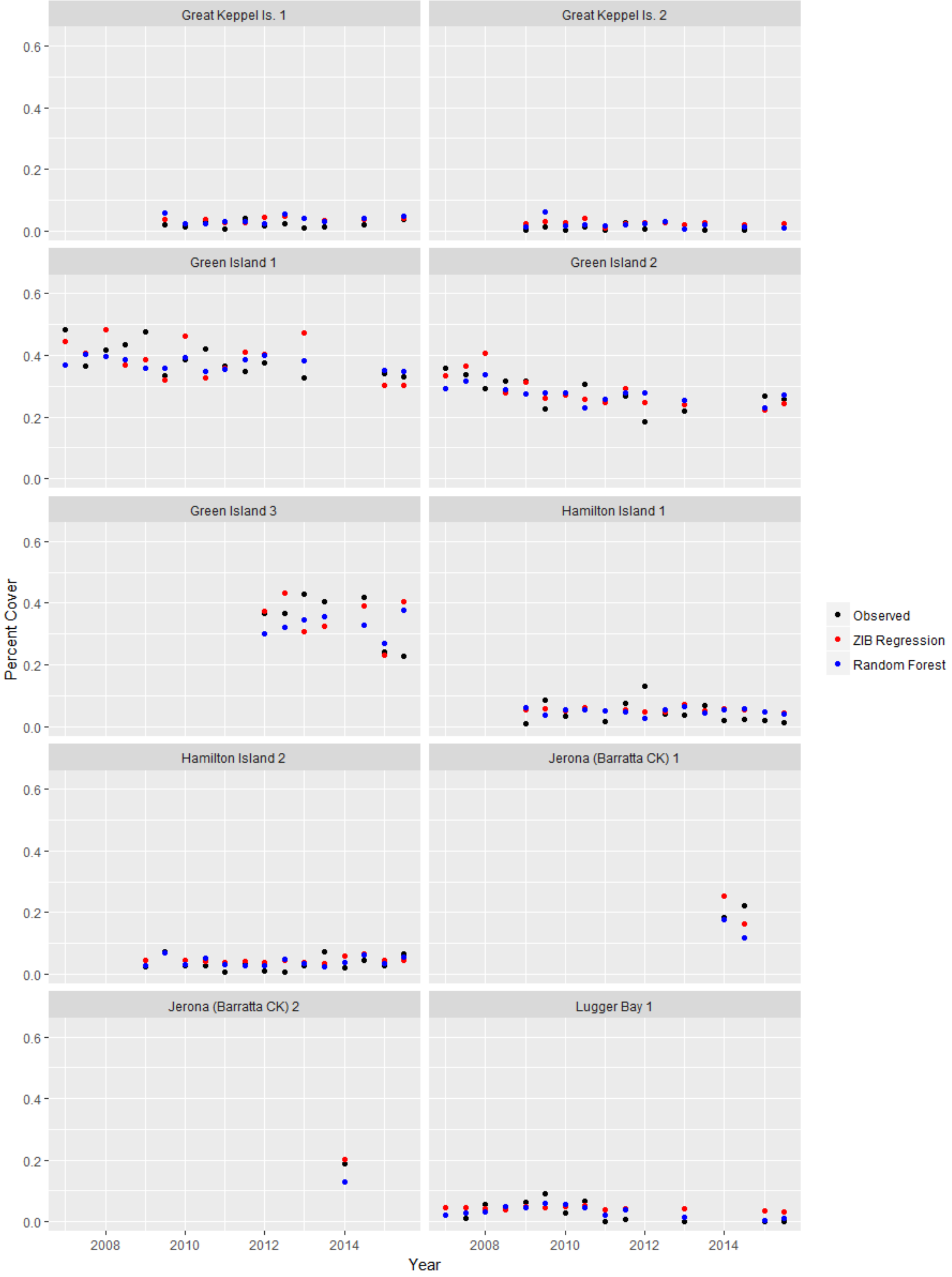

Figure B2: Observed (black) and predicted percent cover of seagrass for the zero inflated beta regression (red) and random forest (blue) for individual sites using model M4.

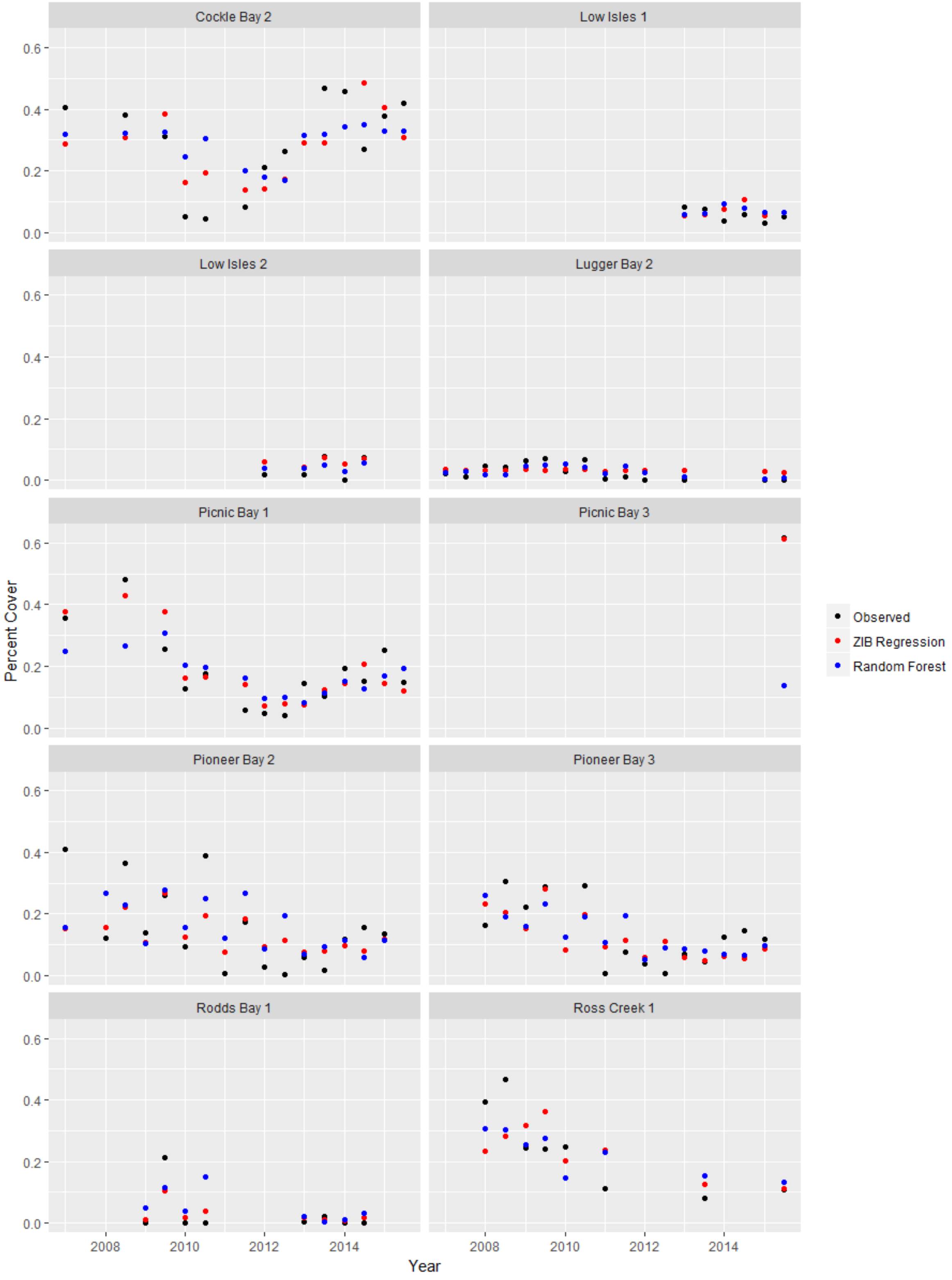

Figure B3: Observed (black) and predicted percent cover of seagrass for the zero inflated beta regression (red) and random forest (blue) for individual sites using model M4.

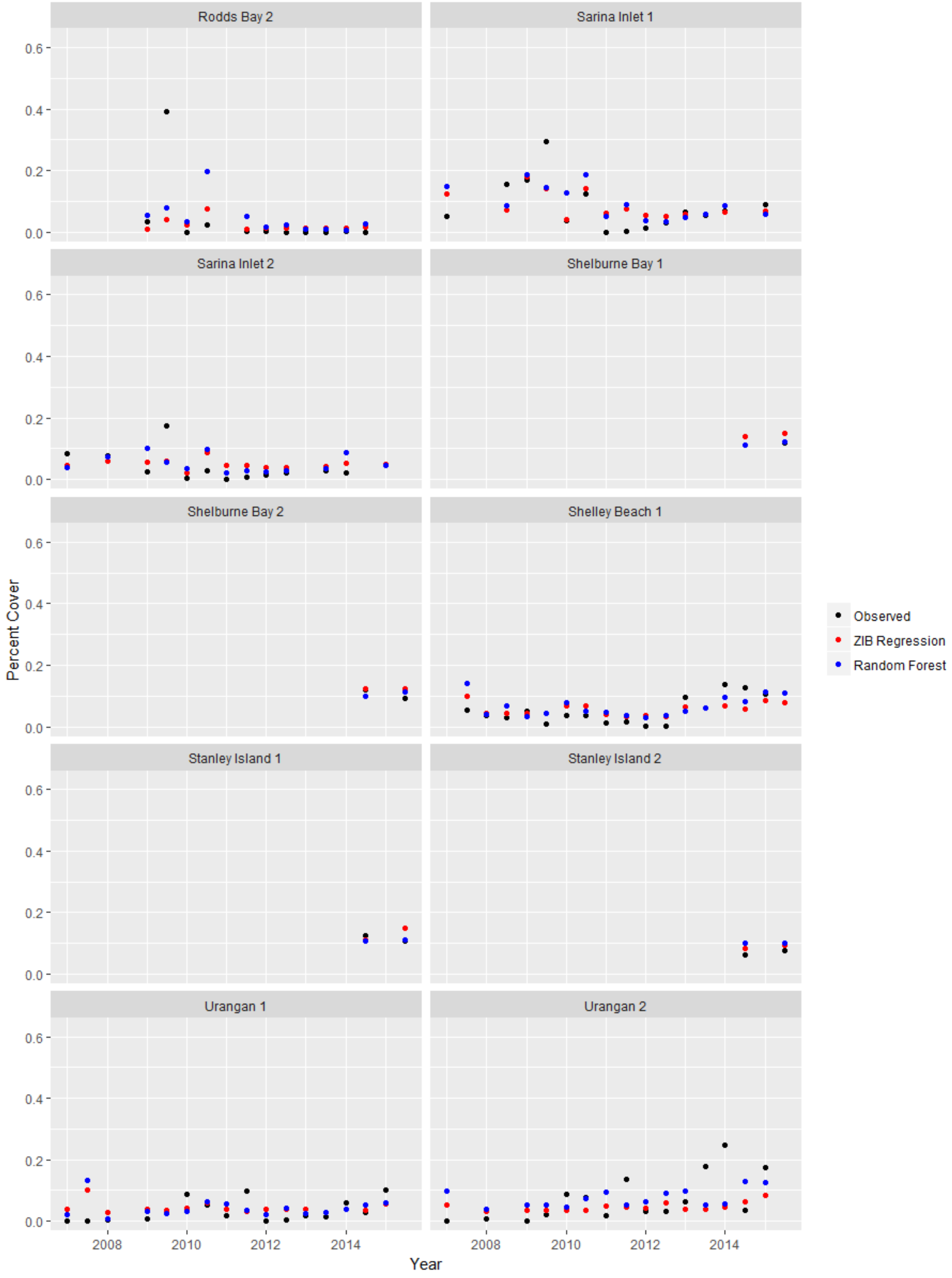

Figure B4: Observed (black) and predicted percent cover of seagrass for the zero inflated beta regression (red) and random forest (blue) for individual sites using model M4.

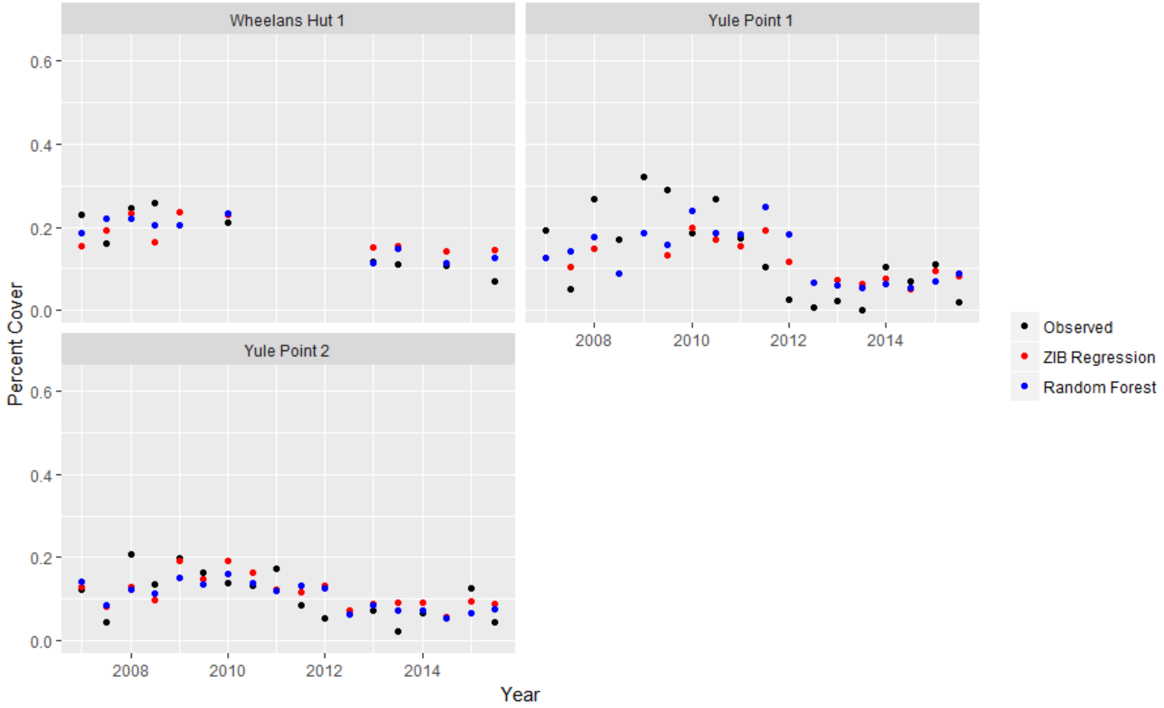

Figure B5: Observed (black) and predicted percent cover of seagrass for the zero inflated beta regression (red) and random forest (blue) for individual sites using model M4.

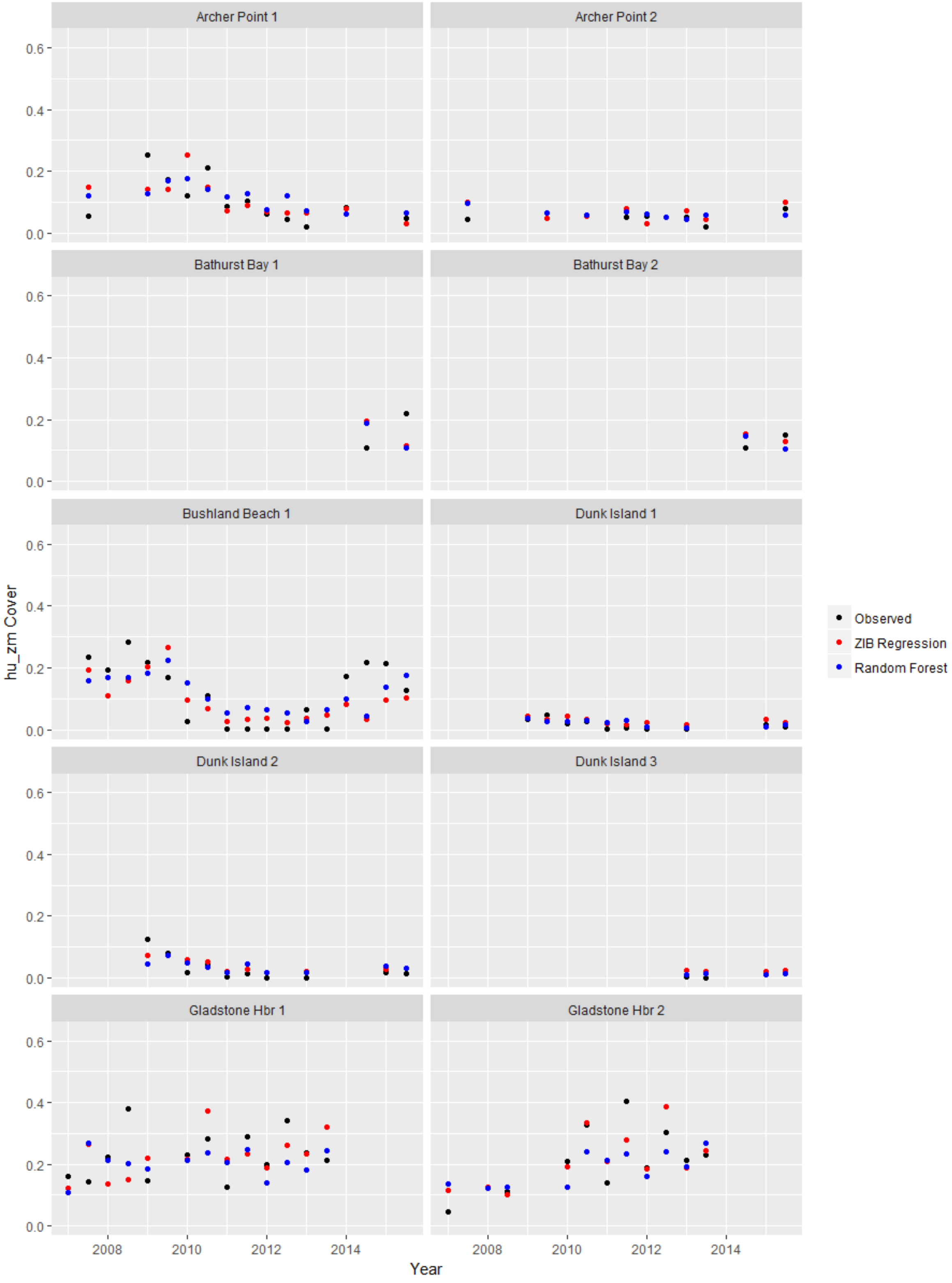

Figure B6: Observed (black) and predicted percent cover of weighed HU and ZM seagrass for the zero inflated beta regression (red) and random forest (blue) for individual sites using model M4.

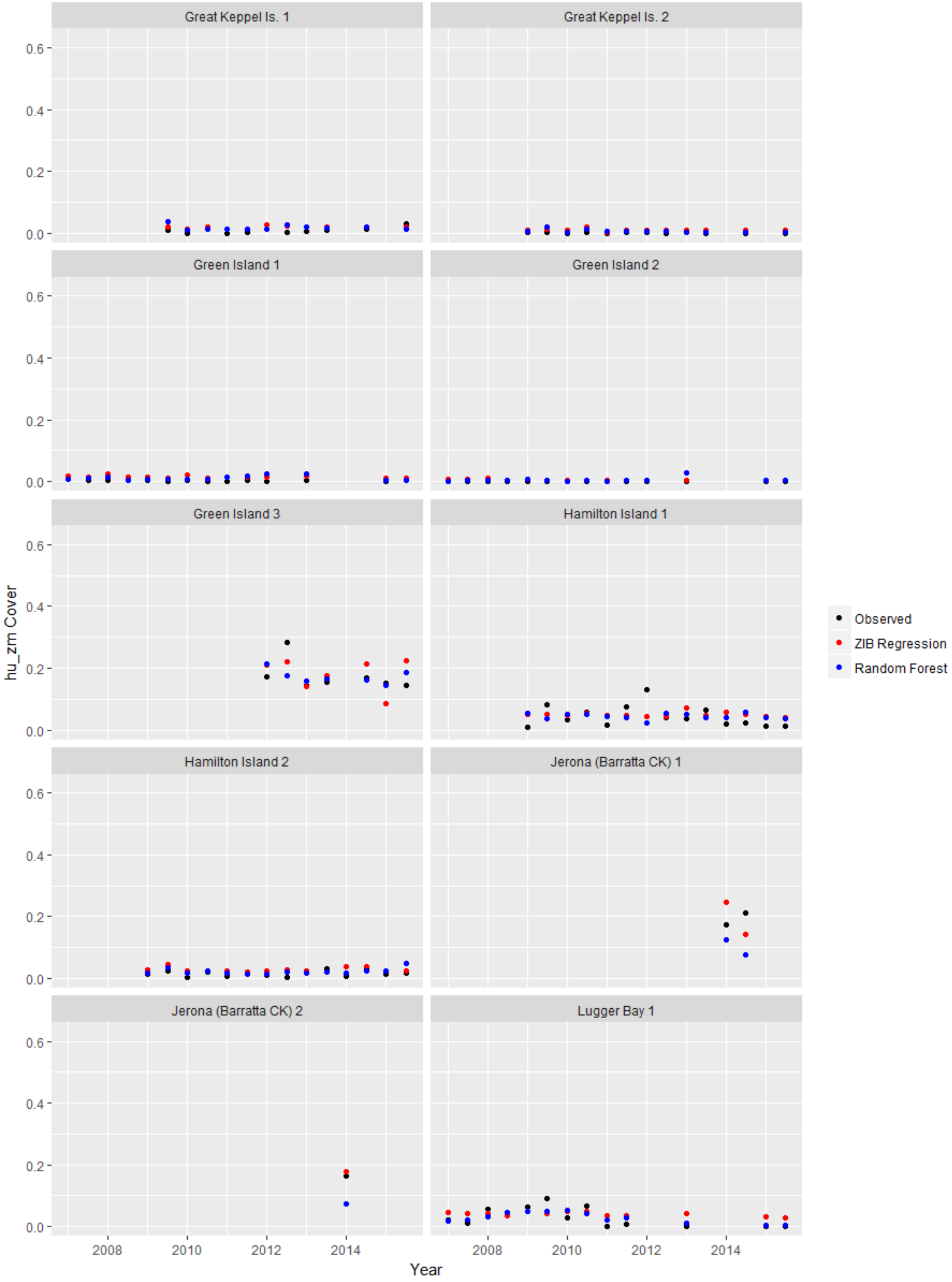

Figure B7: Observed (black) and predicted percent cover of weighed HU and ZM seagrass for the zero inflated beta regression (red) and random forest (blue) for individual sites using model M4.

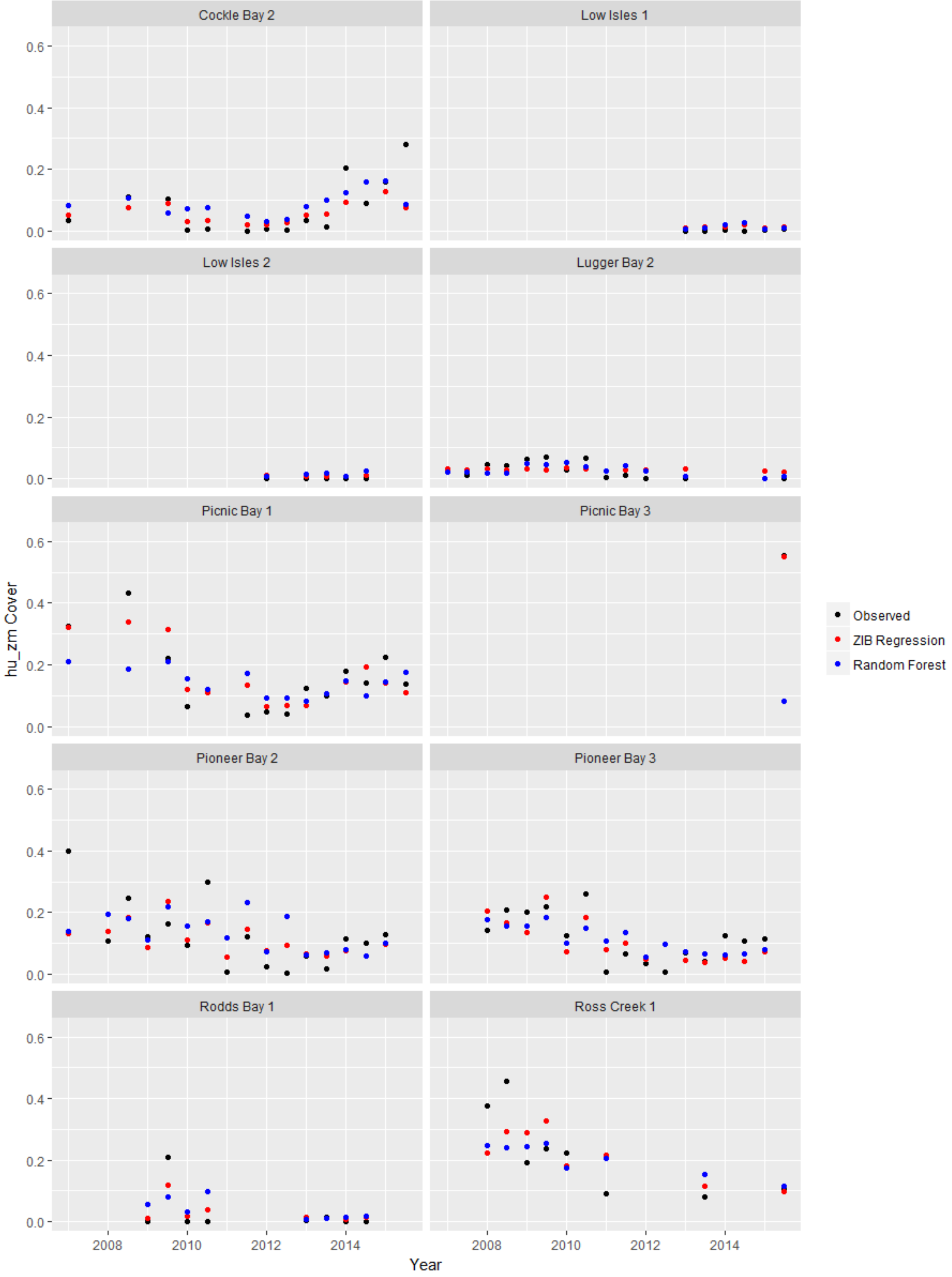

Figure B8: Observed (black) and predicted percent cover of weighed HU and ZM seagrass for the zero inflated beta regression (red) and random forest (blue) for individual sites using model M4.

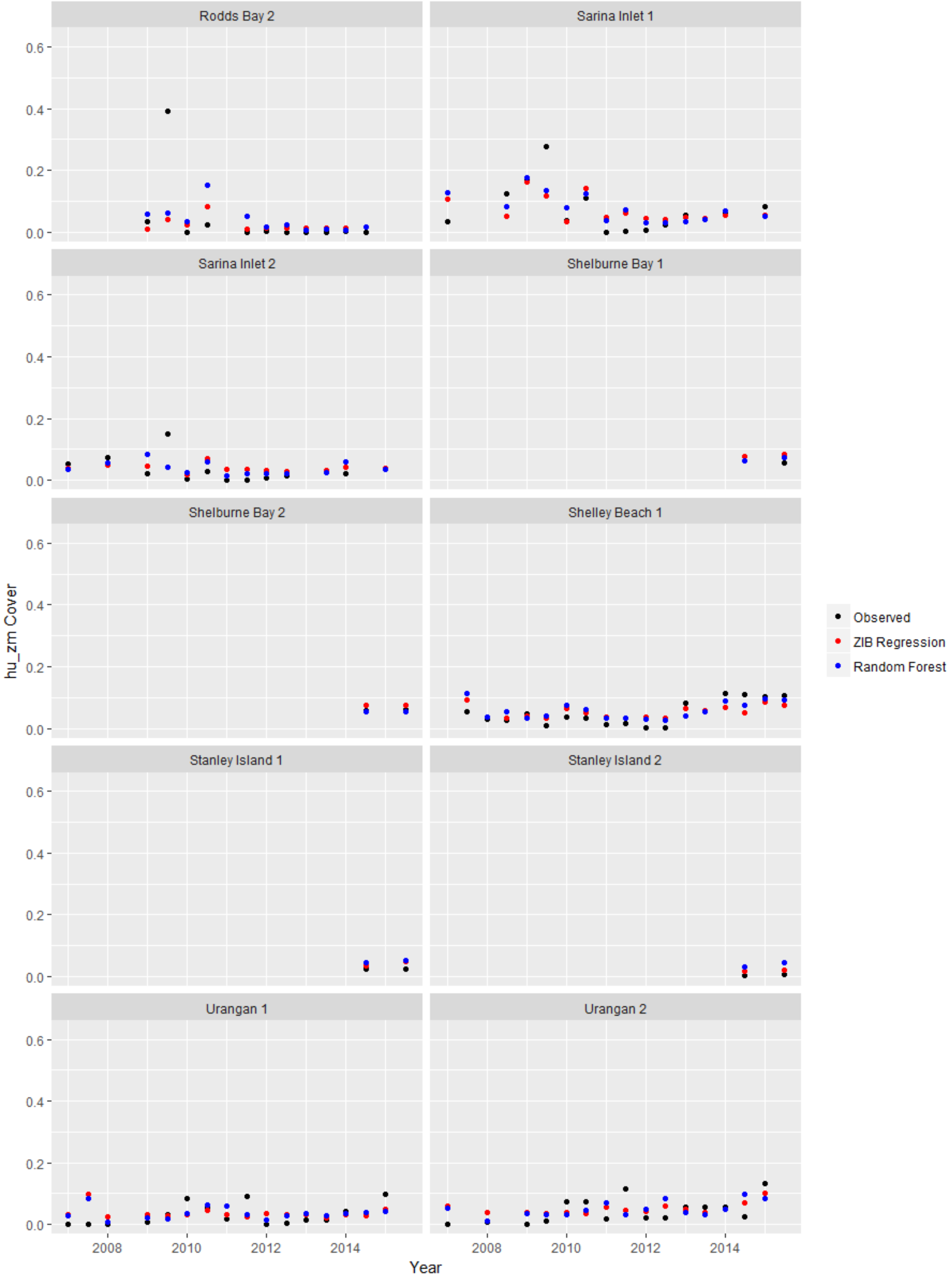

Figure B9: Observed (black) and predicted percent cover of weighed HU and ZM seagrass for the zero inflated beta regression (red) and random forest (blue) for individual sites using model M4.

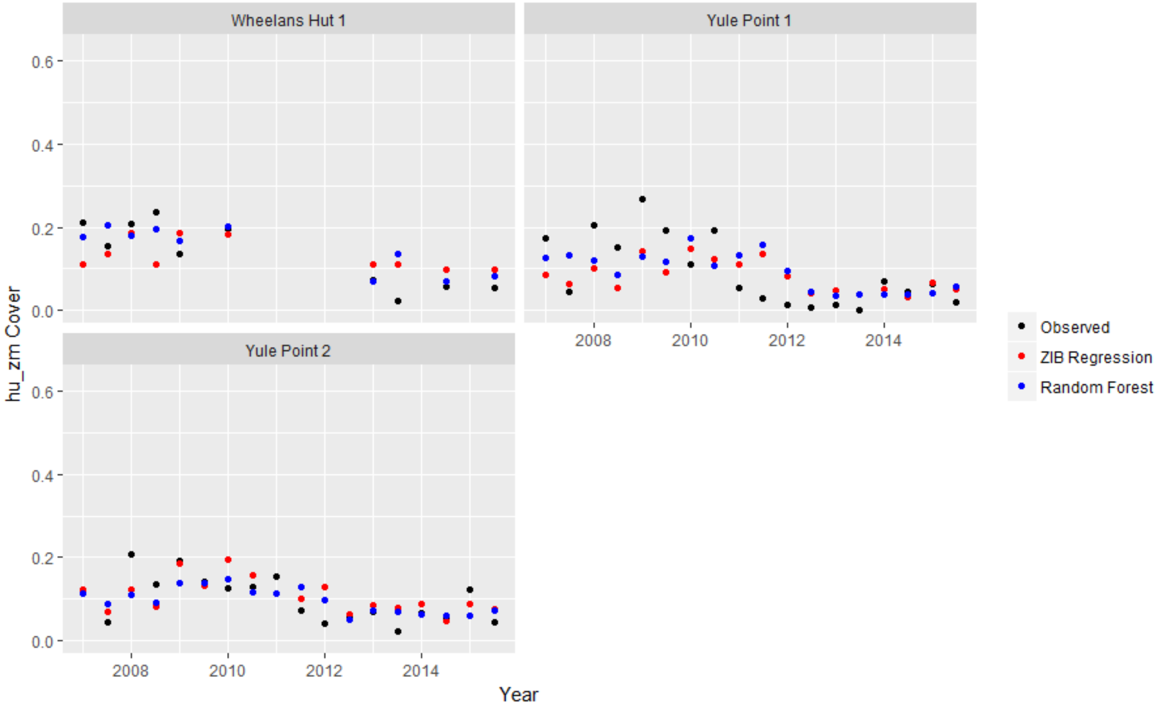

Figure B10: Observed (black) and predicted percent cover of weighed HU and ZM seagrass for the zero inflated beta regression (red) and random forest (blue) for individual sites using model M4.