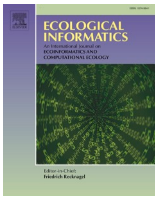
Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf



Synthetic data for reef modelling

Rose Crocker [a,*], Barbara J. Robson [a,b], Chinenye Ani [a], Ken Anthony [a], Takuya Iwanaga [a]

[a] Australian Institute of Marine Science, 1526 Cape Cleveland Rd, Cape Cleveland, QLD 4810, Australia
[b] AIMS@JCU, DB17-148, James Cook University, Townsville, QLD 4811, Australia

ARTICLE INFO

Keywords:
Synthetic data
Reef modelling
Data pipeline
Model testing and validation
Decision support tools
Machine learning
Neural networks
Synthetic data vault

ABSTRACT

Synthetic data mimics the statistical properties of real-world datasets while removing reference to sensitive or confidential information in the original dataset (Quintana, 2020). Synthetic data is also useful for general model testing and development, with many methods available for generating data from machine learning models (Raghunathan, 2021). Although not widely used in the context of ecological and environmental modelling, synthetic data can support and accelerate model testing and analyses where rightsholders are sensitive to data disclosure for study areas, or data collection is expensive.

In the context of reef modelling, synthetic data can be used to support model analyses that can be published without referring to specific sites, reefs, or study areas. This is desirable in the context of decision support for restoration of the Great Barrier Reef. The Reef has many stakeholders and release of early modelling results for intervention scenarios for specific areas would be premature until management or intervention strategy options have been discussed with stakeholders and/or rightsholders. Synthetic data allows a path to publish model and method demonstrations to share knowledge with the reef decision support community without prematurely suggesting policy recommendations for reefs which are sensitive to rightsholders or stakeholders.

We showcase a synthetic data pipeline developed for the reef decision-support system ADRIA (Adaptive Dynamic Reef Intervention Algorithms), using methods from the Python package Synthetic Data Vault (Patki et al., 2016) and others. The synthetic data models are developed to emulate the statistics of case-study reefs for publishing decision-support tool demonstrations, testing and method validation without revealing sensitive reef site information. This pipeline includes developing models for tabular (benthic/compositional reef data), spatial-temporal (wave and heat stress data) and spatial network data (coral larval connectivity). Conditional sampling methods which connect spatial relationships across datasets are used to develop synthetic reef data packages which mimic the statistical properties of the original dataset. The utility of the synthetic data is demonstrated on a sample reef data package, and methods used for anonymizing the data are detailed. The results are discussed in the context of formalizing synthetic data for reef modelling. All synthetic data code is available at ADRIA-synthetic-data/README.md at v0.1.0 · open-AIMS/ADRIA-synthetic-data (github.com), DOI: https://doi.org/10.5281/zenodo.10158323.

1. Why synthetic data for reef modelling?

Here we use the definition of synthetic data as data generated to mimic the statistical properties of real-world datasets while removing reference to sensitive or confidential information contained in the original dataset (Quintana, 2020). Synthetic data allows full exploration of policy-relevant datasets while allaying privacy and sensitivity concerns. This is desirable in the context of decision support for research and management domains, including restoration of the Great Barrier Reef (GBR). Many GBR rightsholders would opt for not revealing early modelling results for intervention scenarios at specific sites until a policy framework has been formally agreed upon. For example, Traditional Owners and tour operators on the Reef may wish to keep the results of exploratory analyses of intervention options and projected outcomes for their areas private until a mutually beneficial strategy has been agreed on. For these reasons, it is desirable to create synthetic datasets which

Abbreviations: GBR, *Great Barrier Reef*; *SDV*, *Synthetic Data Vault*; *GAN*, *Generative Adversarial Network*; *CNN*, *Convolutional Neural Network*; *TVAE*, *Tabular Variational Auto-Encoder*; *PAR*, *Probabilistic Autoregressive.*.

* Corresponding author.
E-mail address: r.crocker@aims.gov.au (R. Crocker).

https://doi.org/10.1016/j.ecoinf.2024.102698
Received 12 December 2023; Received in revised form 18 June 2024; Accepted 19 June 2024
Available online 20 June 2024

sufficiently represent the ecological and environmental conditions of reefs to demonstrate model or decision-tool functionality in publications, but do not reference or disclose actual GBR reef sites.

Further, synthetic data has utility for reef intervention and management research as it can facilitate efficient decision support model testing and validation. Synthetic data inputs can be generated and then augmented such that decision support model outputs are predictable, and any deviation from the expected outputs indicates a bug has been introduced to the code base. This attribute is often capitalized on in the development of artificial intelligence and machine learning models, where there may not be enough real-world data available to evaluate the model validity (Nikolenko, 2021). Examples in the marine modelling space include Watson (2015), where synthetic data sets of sea level rise are built to have predictable mean signals, allowing training sea level rise models despite complex and largely unknown interactions between ocean dynamics and sea level rise. In a similar vein, Wilson et al. (2018) overcome limitations of collected seabed classification data by creating synthetic datasets to support species distribution modelling and more accurate representation of biogeochemistry in marine ecological models. Applications are also common in developing species identification models for automated species mapping, such as for generating larger volumes of marine acoustic trawl surveys (Allken et al., 2019), or underwater imagery for abundance monitoring (Mahmood et al., 2020). In the context of decision tools and models for restoration and management of the GBR, tests of synthetic datasets with known outcomes will improve the robustness of validifying new iterations of the models and decision support tools under development.

Finding and validating a suitable model for generating synthetic data can be time consuming, but once the process is automated it can afford time and efficiency gains in the development process. Generating reef data for modelling can be time consuming and expensive, both due to fieldwork required and/or computationally expensive models. For example, generating coral larval connectivity data for model runs can take days to weeks with the use of a high-performance computer. With a synthetic data generation model, this data can be generated more quickly from old data to create data sets for testing, validation and method/model demonstration in publications.

In this paper we detail a synthetic data pipeline developed for three main applications:

1. To demonstrate the functionality of a reef decision support tool (ADRIA) in publications without the use of potentially sensitive GBR datasets,
2. To develop test data packages for ADRIA's testing suite which contain a small number of sites so that tests run quickly but also cover relevant environmental and ecological conditions,
3. To explore the impact of input data layers on ADRIA's decision support core, investigating key questions about the sensitivity of restoration decision outputs to variability in input data layers.

The intention to create synthetic reef data here is not to guide specific decision or policy recommendations, but for testing, analysis, and validation. For example, to create a small 10-site test set for the reef decision support tool ADRIA, a 10-site data package was generated using the methods developed here. Key decision-influencing properties of these sites, such as the heat stress they experience in bleaching years and capacity for coral cover, were then adjusted up or down relative to the set of sites mean value so that certain sites should always be chosen for intervention activities by ADRIA's site selection algorithms. A series of automated tests on the 10-site test set is now run whenever new changes are introduced to ADRIA.

Specifically, (and as will be detailed in the following section) the model demonstrations, testing and analysis is not focused on an ecological model but on decision support algorithms which use environmental input layers (updated at each time step using an ecological model) as criteria to decide where to implement restoration activities. The data is intended for decision-algorithm validation, testing and publishing demonstrations where real data should not be used due to privacy and sensitivity concerns. Thus, the goal in generating these synthetic data sets is to emulate sensible environmental and ecological dynamics to test, demonstrate and explore these decision support algorithms.

A key challenge in generating synthetic data for reef restoration decision support is that decision support models often require many different input datasets which can be challenging to create a unified synthetic data model for. This challenge is approached here through initially defining a narrow scope for the application of our models (testing and validation) which will be expanded as the models are iteratively improved and key relationships between datasets (which are likely complex) are better understood through specialised research. This paper outlines the first iteration of our synthetic data models, designed to develop data for model testing, validation, and functionality demonstrations. The models developed here may not individually be the best available for the data set they aim to synthesize but were chosen to satisfy a sufficient level of statistical emulation, computational speed, model flexibility and compatibility to generate data packages for the purposes listed above. The methods used to demonstrate the utility of the data models (often called general utility) (Snoke et al., 2018), were chosen for their applicability to a range of data types (allowing comparison across different input datasets) and high compatibility with the Synthetic Data Vault python package. These are discussed in further detail in Section V.

2. Reef decision support model – ADRIA

We develop synthetic input datasets for the reef modelling and decision-support tool, the Adaptive Dynamic Reef Intervention Algorithm (ADRIA). ADRIA is a decision-support tool designed to assist in the planning and implementation of restoration projects on the Great Barrier Reef (Iwanaga et al., 2024). ADRIA is informed by a parsimonious mathematical model of coral survival, fecundity, growth, and density-dependent recruitment of six species with six size classes. The coral model is forced by environmental drivers, ecological processes, and state variables, and restorative and/or protective interventions. The decision-support algorithms in ADRIA use these environmental and ecological data layers as heuristic criteria to dynamically select locations best suited to implement restoration activities under different ecological and environmental conditions.

ADRIA's ecological model and location-selection algorithms require a suite of ecological and environmental input data layers, as detailed in Fig. 1. To inform spatial planning in ADRIA, the reef location of interest is divided into a series of polygons based on benthic composition, which are used to designate 'sites' as locations at the centroids of these polygons. The process of site selection for a particular decision instance in ADRIA is illustrated in Fig. 2. First a series of data layers are selected based on relevance to the restoration decision. These layers are aggregated into criteria, so that there is a single value for each reef site for each criteria (a decision matrix). Examples of criteria include the mean heat stress a site experiences over the 5 years following intervention, the mean site depth, or the area available for coral growth at the site. The criteria are normalised and then can be used to filter sites based on thresholds, such as those too hot or deep to implement a restoration activity. Finally, the criteria are weighted according to their importance or the decision maker's preference for a particular restorative intervention and aggregated to give a single value for each site. The sites are then ranked from highest value to lowest and the top N sites are chosen to implement the restorative activity at.

The synthetic data sets are developed to test, demonstrate, and explore ADRIA's core decision support algorithms. The goal in creating the synthetic input data sets is to allow demonstrations and automated testing of ADRIA's decision support algorithms as criteria are added and other changes introduced, as well as to explore the sensitivity of the

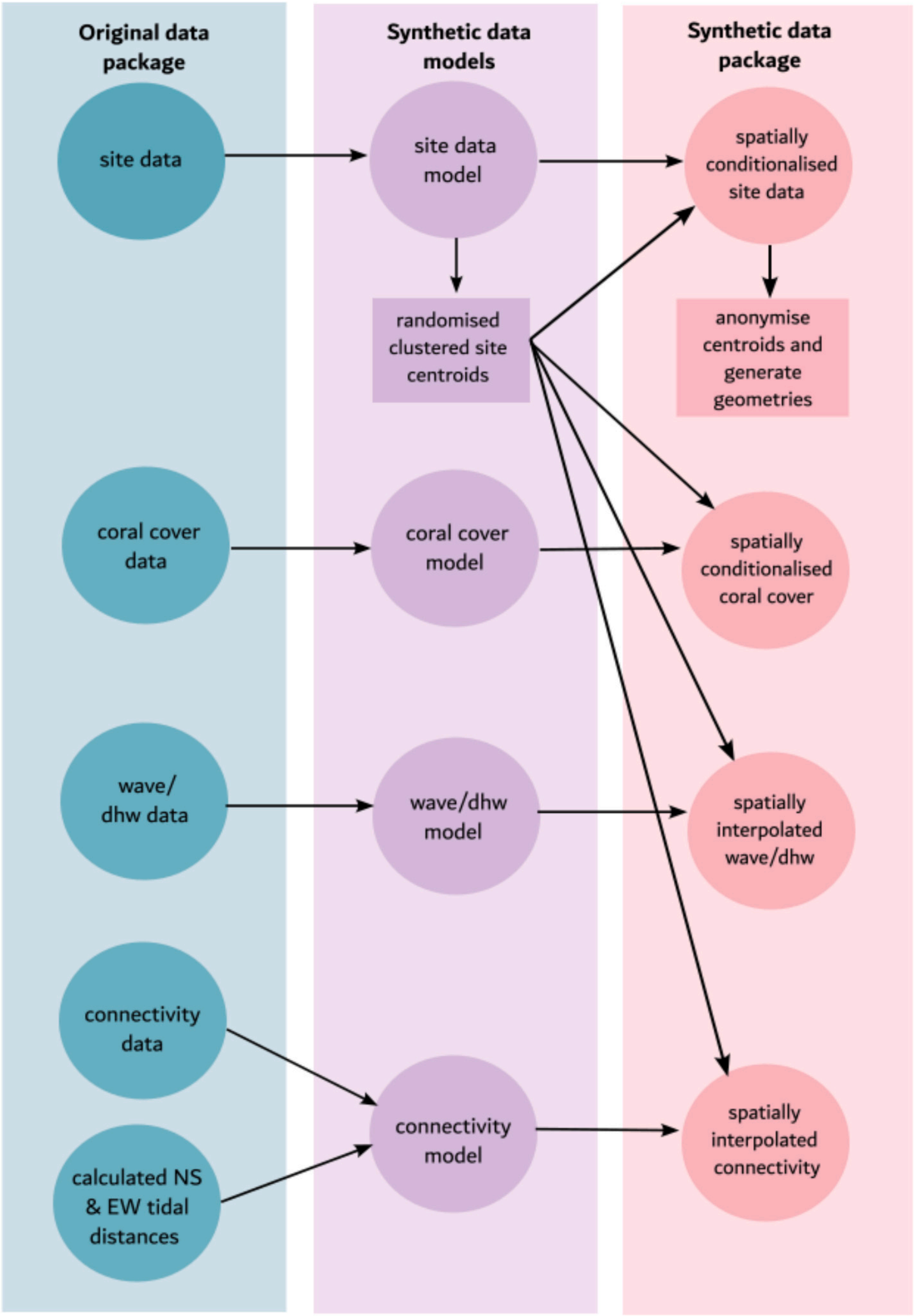

Original data package
Synthetic data models
Synthetic data package
site data
site data model
spatially conditionalised site data
randomised clustered site centroids
anonymise centroids and generate geometries
coral cover data
coral cover model
spatially conditionalised coral cover
wave/ dhw data
wave/dhw model
spatially interpolated wave/dhw
connectivity data
calculated NS & EW tidal distances
connectivity model
spatially interpolated connectivity

Fig. 1. The synthetic data pipeline used to create synthetic data packages for ADRIA. Synthetic data models are created for each of the datasets and conditionalised on the synthetic site data to create a spatially consistent synthetic site data package.

rankings of sites to variations in different input data layers. The key data layers which become criteria and are synthesized here are detailed in the following paragraphs.

The key spatial data layer for ADRIA is the ‘site data’, which details the latitudinal and longitudinal positions of the polygon centroids, the polygon areas, their maximum capacity for coral as a percentage of the area (k), categorical benthic composition, and other key identifying information. This is used for several key criteria during site selection, including mean site depth, site coral cover area and protection zoning category. Another data layer specifies the initial coral cover at each site for each of the 6 coral species and 6 size classes (36 categories), which is used to initialize ADRIA’s coral growth model. This, along with site area and carrying capacity from the site data, forms a criterion detailing the area of space available for coral to be planted (a restorative intervention), and another criteria describing the current area of coral which can be protected using solar radiation management (another restorative intervention).

Key environmental data layers include time series data which specify the mean heat and wave stress that each of the reef sites are projected to experience over a time span of 100 years. These are used in site selection criteria which detail the average heat and wave stress a site will experience over some duration of time following a restorative intervention, with the site selection algorithm seeking to avoid high stress sites. Heat stress, expressed as degree heating weeks data, and wave stress, expressed as mean wave height in metres, are three-dimensional data cubes, with dimensions year, climate replicate and site.

Connectivity data, which describes the larval connectivity for a given year between each site in the region of interest, is represented as an *N* by

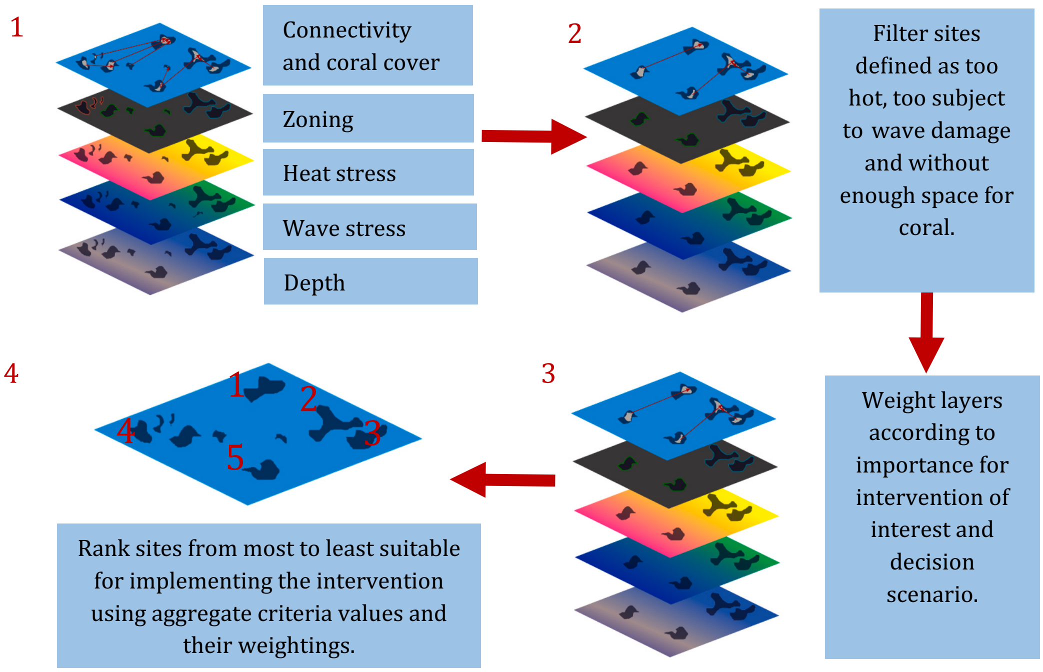

1
Connectivity and coral cover
Zoning
Heat stress
Wave stress
Depth
2
Filter sites defined as too hot, too subject to wave damage and without enough space for coral.
3
Weight layers according to importance for intervention of interest and decision scenario.
4
1
2
3
4
5
Rank sites from most to least suitable for implementing the intervention using aggregate criteria values and their weightings.

Fig. 2. The restoration location selection process in ADRIA (criteria illustrated are examples from a larger selection).

N matrix, where entry (i,j) is the connectivity between sites i and j. The connectivity data is used in the site selection algorithms to form criteria which favour sites providing larvae to sites experiencing lower heat and wave stress. The models and assumptions used to create these original data layers are detailed in Appendix A and the connectivity and DHW data are discussed in more detail in Cresswell et al. (2023).

3. Synthetic data generation

As mentioned, we use the definition of synthetic data as data generated to mimic the statistical properties of real-world datasets while removing reference to sensitive or confidential information contained in the original dataset (Quintana, 2020). In time series analysis, such data may also be referred to as “surrogate data”, which reproduces statistical properties of an original dataset, such as autocorrelation structure. Most importantly, such data is generated “artificially”, by a model and/or process to emulate the statistical properties of another dataset.

Use of synthetic data began as a means to allow sharing of confidential census data. It has gained popularity in public health, pharmaceutical and medical research due to demand for retaining data privacy while releasing important study results or demonstrating new analysis methods (Chen et al., 2021). More recently it has become common in machine learning research due to demand for large training data sets where data is limited (James et al., 2021). Synthetic data is also useful for general model testing and development, with methods available for generating data from models ranging from simple to complex (Raghunathan, 2021).

Synthetic data can be generated using a variety of methods including statistical methods, process-based modelling, classical machine learning methods and deep learning methods. The first synthetic data methods developed used techniques from Statistical Disclosure Control, which designates transformations of either some or all columns of a data set to obscure private and sensitive details. These techniques were then coupled with frequentist approaches such as confidence intervals to check the transformed data reasonably retained the statistical properties of the original data (R.J.A, Liu and Raghunathan, 2004). In a similar vein, statistical permutation methods can be used to create synthetic data sets by calculating a test statistic for the original dataset and randomly permuting the data points. The test statistic is then recalculated in the permutated sets and compared to the original to see if it falls within the null distribution (Berry et al., 2018). This can also be applied as Monte Carlo resampling, where many datasets are simulated via reshuffling the original data before testing the null hypothesis that the distribution of the shuffled data is the same as that of the original dataset (Ernst, 2004). Bayesian approaches treat the unsampled proportion of the original data as missing information and construct a posterior distribution for the synthetic data conditional on the observed data (Raghunathan, 2021).

Synthetic data can also be generated by creating process-based or surrogate models for the original data. Process-based models use known properties of the system (e.g. physical, chemical, biological processes) to generate data which can then be compared to collected data via statistical methods. Developing process-based synthetic data models, however, can be time consuming, models can be slow to run and can be highly sensitive to input parameters. This is certainly the case with many reef data sets which can be high in uncertainty, highly parameter sensitive and computationally expensive to model. The coral larval connectivity data discussed in later sections, for example, can take many days to weeks to generate depending on the spatial domain. Surrogate models are designed to mimic the behaviour of a complex process-based model while being computationally cheaper to run. The surrogate model is generally developed based on a selection of input data points and the complex model’s response to those data points. Surrogate models can be highly effective, but the data points used to build the model on must be sampled carefully and the surrogate model’s parameters need to be optimized to balance variance and bias. Surrogate models can be developed using reduced order approximations, such as using radial basis functions, through multi-fidelity models such as Kriging, which combine data sources of differing fidelity, or via machine learning methods (Alizadeh et al., 2020).

Classical and deep learning methods are increasingly being used to generate synthetic data as they can perform better at emulating complex patterns and relationships in the original datasets, and do not require knowledge of prior or posterior distributions for dataset variables. Classical machine learning methods include regression, K-nearest neighbours, and support vector machines, and can perform better on smaller data sets and take less time to train than deep learning methods. Deep learning methods are based on neural networks, including convolutional neural networks (CNN), auto encoders and generative adversarial networks (GAN), and are better at learning sophisticated patterns in data but can be more computationally expensive and require more data to train (Endres et al., 2022).

Synthetic data is not widely used in conservation and environmental management, but its benefits are being increasingly recognized and capitalized upon. Watts et al. (2011) use synthetic satellite data to improve land-classification models for assessing carbon sequestration and soil quality, due to lack of cloud-free satellite data restricting model accuracy. Wimmer and Finger (2023) demonstrate the efficacy of synthetic data methods in replicating agricultural production data, recommending the use of synthetic data in agricultural economics studies, where data often has policy implications and cannot be published for confidentiality reasons. Fassnacht et al. (2018) use synthetic data to evaluate estimates of forest biomass, citing lack of available test data for fine-tuning and assessing method robustness. Reviews of data sharing in ecology have stressed the need for better access to important data sets to fully capitalize on the wealth of data being generated (Reichman et al., 2011). Synthetic data offers a means to address essential environmental and ecological questions where collection and/or integration of additional real-world data is prohibitively expensive, existing data cannot be made publicly available, or data collection has faced policy and funding challenges (Poisot et al., 2016).

4. Synthetic data utility/validation measures

A range of metrics were used to evaluate the statistical validity of the synthetic data models developed here. The metrics are computed using the SDMetrics library (DataCebo, 2023) and Scikit-learn, and were chosen to represent a range of key properties when generating synthetic data, including statistical similarity, coverage and uniqueness. Statistical similarity metrics evaluate how well the synthetic data captures the distribution and correlations between variables in the original data. Coverage metrics evaluate how well the synthetic data captures the incidence of minimum and maximum values observed in the original data. Uniqueness metrics evaluate the uniqueness of the original and synthetic datasets to assure the synthetic data is not just a complete copy of the original, which can happen if model mode collapse occurs (Zhang, 2021). The metrics chosen allow comparisons of success across the models used, and can be disaggregated into scores for individual variables, allowing understanding of which variables are best represented by the models. Along with these measures, in section VII, ADRIA model outputs are compared for inputs of original and synthetic data to assess the suitability of the data in representing the range of outcomes represented by the original dataset. As the main focus in creating the synthetic data sets is to test and demonstrate the reef site selection algorithms in ADRIA, statistical similarity was considered the primary requirement, as the decision algorithms use statistical aggregates of data layers as criteria to determine restoration decisions. The main metrics used, and descriptions of their calculation are summarised in Table 1 below.

5. Data generation methods

As the input datasets represent different data types (tabular, timeseries, relational), different synthetic data models were selected for each of the input types, and statistically validated using the utility measures described in the previous section (DataCebo, 2023). Appending geo-coordinates (latitudinal and longitudinal data) to the coral cover, DHW, wave and connectivity datasets during model training allow the use of conditional sampling to produce synthetic data packages for ADRIA that are spatially consistent with the synthesized site data. The synthetic latitudes and longitudes are then anonymised to remove any references to real reef sites and provide anonymity. This is done by a) sampling the sites in a new, user-controlled spatial configuration and b) shifting the entire set of sites a random distance along the latitudinal and longitudinal directions. The synthetic data package pipeline which has been

Table 1
: Data utility metrics descriptions.

Metric	Description
χ^2test	For categorical variables in the datasets. Tests the hypothesis that the synthetic columns and original data columns are from the same distribution. Here the inverse value is used so that a high value (between 0 and 1) indicates the distributions are not different in a statistically significant way.
Kolmogorov-Smirnov test	For continuous variables in the datasets. Tests the overlap of the synthetic and original data distributions for each variable. The null hypothesis of the test is that a particular variable from the original and synthetic data follow the same distribution. The cumulative distributions of the original and synthetic data variable are estimated and if the mean distance between the distributions is small enough the null hypothesis cannot be rejected. Here the inverse value is used so that a high value (between 0 and 1) indicates the distributions are not different in a statistically significant way.
Correlations scores	For pairs of continuous variables the correlation similarity (0–1) tests the measures the degree to which the two variables are correlated in the same way in the original and synthetic data sets. High scores imply the pairwise Spearman correlation coefficient are highly similar, while low scores imply the are very different. The score is calculated as: $score = 1 - 0.5\lvert S_{A,B} - R_{A,B}\rvert$ Where $S_{A,B}$ is the correlation between columns A and B in the synthetic data and $R_{A,B}$ is the correlation between columns A and B in the real data. For pairs categorical variables or pairs of categorical and continuous variables, the contingency similarity is computed using the Total Variation Difference: $TVD = 1 - 0.5\sum_{\alpha\in A}\sum_{\beta\in B}\lvert S_{\alpha,\beta} - R_{\alpha,\beta}\rvert$ Where α are the categories in column A and β are the categories in column B and R and S are the real and synthetic frequencies for those categories. For pairs of continuous and categorical variables, the continuous variable is discretised via binning to calculate the value.
Mean data quality score	A mean score combining the χ^2 test for categorical variables, K—S test for continuous variables and pair-wise correlation score between variables across the synthetic and original datasets. The higher the percentage (between 0 and 100) the better the synthetic data statistically emulates the original dataset. For example, consider a dataset containing one categorical variable with inverse χ^2 test score 0.7 and two continuous variables with inverse K—S test scores 0.8 and 0.9 respectively. The correlation score between the two continuous variables is 0.85, and between the continuous and categorical are 0.86 and 0.89. This would give a quality score of $\left(\frac{statistical\ similarity\ mean}{2} + \frac{correlations\ mean}{2}\right) \times 100 = \left(\frac{0.7+0.8+0.9}{2*3} + \frac{0.85+0.86+0.89}{2*3}\right) \times 100 = 83.3\%$
Mean data diagnostics scores	A set of scores assessing the coverage of the data (how well does it cover the range of values in the original data?) and how unique the dataset is (i.e. is it an exact copy of the original or just statistically similar?). The coverage score (0–1) assesses how well the synthetic data covers the range of values in the original dataset. For continuous variables this is calculated as $score = 1 - \left[max\left(\frac{min(s) - min(r)}{max(r) - min(r)}\right) + max\left(\frac{max(r) - max(s)}{max(r) - min(r)}\right)\right]$ Where r is a dimension of the real data and s the same dimension of the synthetic data. The score is then averaged across comparable dimensions in the two datasets to give the mean coverage score. The boundaries score (0–1) assesses how well the maximum and minimum values match those of the original dataset. It is computed by calculating the min and max values of the original data and then calculating the frequency of values in the synthetic data which are outside this range. A value of 1.0 implies no synthetic data is outside the range of the original data. The synthesis score (0–1) assesses the uniqueness of the dataset compared to the original. It is calculated as $score = 1 - \frac{no.of\ synthetic\ data\ rows\ within\ 1\%of\ the\ original}{total\ no.of\ synthetic\ data\ rows}$

(*continued on next page*)

Table 1 (*continued*)

Metric	Description
Principal Components comparison	Compares the principal components of the original and synthetic data from a Principal Component Analysis (PCA). Principal components are linearly uncorrelated basis vectors for the data set which can be ordered according to those which describe the most variance in the dataset to the least. As synthetic data sets are often very high dimensional, calculating a PCA for the original and synthetic datasets and plotting the first two components on the same plot can give a visual indication (to complement other numerical measures of utility) of how well the original and synthetic distributions overlap. The first 2 components are generally used as these can be easily plotted.
Correlation Matrix	Correlation matrices capture the pairwise correlations between variables in the synthetic and original datasets. Comparing correlation matrices can visualise how well the synthetic data captures relationships between variables in the original dataset. Numerically, the degree to which the synthetic data captures correlations in the original data set is captured in the "Column pairs score" (see *Key variable scores* column in Table 2).

developed for ADRIA is illustrated in Fig. 2 and described in the following sections.

To choose the best models for each data type, several suitable models within the SDV package were compared in terms of their synthetic data quality scores (See section V) and runtime. In the case of the connectivity data, only one model (Gaussian Copula) had reasonable runtimes, so another model (a Generative Adversarial Network) from the y-data-synthetic package (YData, 2023) was also tested. In most cases, the models which gave the highest data quality scores with the shortest generation times were chosen. In some cases, however, where quality scores for two models were close, the scores for individual variables were used to select between models. For example, for the initial coral cover data the Gaussian Copula model and TVAE models performed similarly in their quality scores but on disaggregating the quality scores into variable-specific scores, the TVAE model performs better at predicting the variable \`cover\` while the Gaussian Copula model performs better at predicting the variable \`species\`. Both models were retained as options in ADRIA's synthetic data model repository as either model may be desirable depending on the usage, with the Gaussian Copula model offering faster runtimes and better representation of species relationships, while the TVAE model offers reasonable but longer runtimes and a better representation of cover. Similarly, the GAN model performs better than the Gaussian Copula model for generating connectivity data but is significantly slower. The Gaussian Copula model was also retained as an option alongside the GAN model in the synthetic data repository as it's quality scores are still reasonable. The key results and times from these model comparisons are summarised in Appendix D. In the case study, the TVAE and GAN models are demonstrated for the coral cover and connectivity data respectively.

5.1. Site data

The site data model was developed using a Gaussian Copula model, as this performed best in terms of utility metrics (detailed in Table 2) and runtime. The model uses Gaussian copulas to fit the columns of the dataset to find marginal distributions and then the covariance of each pair of columns to learn the joint distribution. Details are available in (Patki et al., 2016).

One issue with the model, however, was that it did not simulate the spatial clustering of reef sites which generally occurs in real reefs, as shown in Fig. 3. To amend the unrealistically spatially scattered sites generated by the original Gaussian Copula model, M sites are randomly selected from the synthetic data set and N site latitude and longitude positions are generated in randomized, constrained radii around these M sites. The model allows flexibility around the number of nodes to generate sites around and the maximum length of the randomized radii,

Table 2
: Data utility scores for synthetic datasets developed in the case study.

Dataset	Quality score	Key variable scores		Diagnostics score
Site data	*Column Shapes:* 88.81% *Column Pair Trends:* 90.97% *Overall Quality Score:* 89.74%	*Column shapes:*		*Synthesis:* 1.0 *Coverage:* 0.99 *Boundaries:* 1.0
		Latitude	91.5%	
		Longitude	92.3%	
		Site carrying capacity	90.8%	
		Median depth	87.9%	
		Site area	92.4%	
		Column pair trends:		
		Area and site carrying capacity	94.3%	
		Site carrying capacity and median depth	99.0%	
		Area and median depth	98.6%	
		Carrying capacity and longitude	97.0%	
		Carrying capacity and latitude	96.3%	
		Area and latitude	97.3%	
		Area and longitude	97.8%	
Cover data	*Column Shapes:* 90.94% *Column Pair Trends:* 86.55% *Overall Quality Score:* *89.62%*	*Column Shapes:*		*Synthesis:* 0.86 *Coverage:* 0.94 *Boundaries:* 1.0
		Cover	93.1%	
		Species	92.3%	
		Latitude	95.7%	
		Longitude	92.3%	
		Column pair trends:		
		Cover and latitude	98.1%	
		Cover and longitude	93.2%	
		Cover and species	82.6%	
DHW data	*Column Shapes:* 95.34% *Column Pair Trends:* 89.97% *Overall* Quality *Score:* 92.66%	*Column shapes*:		Synthesis: 0.67 Coverage:0.99 Boundaries:0.99
		DHW	93.0%	
		Latitude	92.4%	
		Longitude	95.9%	
		Column pair trends:		
		DHW and year	98.9%	
		DHW and latitude	85.0%	
		DHW and longitude	86.0%	
Wave data	*Column Shapes:* 87.34% *Column Pair Trends:* 86.71% *Overall Quality Score:* 87.03%	Column shapes:		Synthesis: 0.88 Coverage: 0.94 Boundaries: 0.96
		Ub	71.1%	
		Latitude	91.5%	
		Longitude	86.7%	
		Column pair trends:		
		Ub and year	95.5%	
		Ub and Latitude	75.2%	
		Ub and Longitude	90.0%	
Connectivity data	*Column Shapes:* 98.2% *Column Pair*	*Column shapes:* Minimum score 94%, maximum 99.9% (there are as many columns as		*Coverage:* 0.93 *Boundaries:* 0.89 (Synthesis is too computationally time

(continued on next page)

Table 2 (*continued*)

Dataset	Quality score	Key variable scores	Diagnostics score
	Trends: 99.9% *Overall Quality Score:* 99.0%	sites so not all can be reported here).	consuming to calculate for 100 > rows)

so that the user has control over what the clustering of reef sites look like. As previously mentioned, the intention for the data is testing, exploration and demonstration of ADRIA's decision support algorithms. Degree of clustering is an important characteristic for ADRIA's site selection algorithms, as the algorithms can be set to use a criterion based on how tightly clustered sites selected for restoration are to reduce the risk that all selected sites are impacted by a localised disturbance simultaneously. The criterion is set by specifying how many of the sites selected for restoration are allowed within a single localised cluster, where clusters are defined by a k-means clustering algorithm based on a distances matrix. Consequently, although future work may focus on capturing relationships between benthic properties and site clustering, for this iteration of the data models this approach was sufficient to capture the spatial clustering criteria for testing. It is of interest to better understand what drives these patterns in future investigations, but this simple method achieved the spatial clustering necessary for testing and demonstrating ADRIA's decision support core.

The Gaussian Copula site data model is then conditionalized on the radially sampled latitudes and longitudes to generate a site data set for the clustered site positions while retaining the learnt statistics of the original data set. Geometries are generated for the sites by drawing sufficiently large circles centred on the sampled site latitudes and longitudes to agree with the generated synthetic site areas.

5.2. Initial coral cover data

The initial coral cover model was developed using a TVAE model from SDV, which is based on a Variational Autoencoder model. In this model, an encoder maps the original data to distributions in a lower dimensional latent space. Data is then sampled from the latent space and transformed back to the original space using a decoder. In the learning phase the decoder learns by finding a transformation with minimal loss between the original and final data distributions. This loss is quantified using the evidence lower bound (ELBO), which transforms intractable inference problems into optimisation problems which can be solved using gradient methods (Xu et al., 2019).

The model learns the spatially dependent distribution of cover for each of the six species of coral modelled in ADRIA. The radially sampled latitudes and longitudes from the synthetic site data set are then used to conditionalize the sampling of the synthetic coral cover model so that the synthetic data sets are spatially consistent. Finally, the conditionally sampled cover is distributed over the 6 size classes used in ADRIA for each species according to the mean proportions of cover in each size class for each species in the original data set. The model can also be set up to learn the size class distributions for each species, but was found to become time consuming during the spatially conditioned sampling and does not significantly add to the quality scores of the model.

5.3. DHW and wave data

The synthetic degree heating weeks and wave height data was generated using a Probabilistic Autoregressive model (PAR) from SDV. Auto-regressive models express future time states as a linear combination of states at the previous time step and parameters as coefficients, plus a time-dependent error term (Zhang et al., 2022). The PAR model allows designation of 'entity' and 'context' variables, which are variables in the dataset which are constant with time. In the case of the DHW and wave data sets for ADRIA, the entity variable was set as the site ID and the context variable (which is an additional variable contextualizing the entity variable) was set as the latitude and longitude corresponding to the site ID.

The DHW and wave data is also conditionalized on the synthetic site data latitudes and longitudes when sampling the final synthetic datasets. The original wave and DHW datasets contain a 'climate replicate' dimension, representing statistical realisations of possible climate futures. To represent this in the synthesized data, climate replicates were sampled by randomizing the original replicate the data model was learnt from and then conditionally sampling the time series individually for a specified number of replicates. The user chooses how many replicates to sample and how many samples there will be in the final dataset. For example, if 5 replicates were sampled with a final dataset of 50 replicates, a model will be trained for 5 replicates and 10 samples will be drawn for each replicate for a final set of 50 replicates. The more replicates which are sampled the longer that training process takes, so the user has flexibility in choosing a number of replicates suitable for time constraints while representing climate variability in the data.

5.4. Connectivity data

The synthetic connectivity data was generated using a Generative Adversarial Network (GAN) model based on the y-data-synthetic Python package (YData, 2023), as this model was able to replicate the sparseness of the original connectivity matrices best and performed best in terms of the utility metrics. GAN models feature a generator component, which learns the latent features of the data to generate a data sample, and a discriminator, which is a classification model which learns to classify real and synthetic data. Backpropagation from the discriminator updates the model parameters in the generator with the magnitude of the update depending on the success of the discriminator in classifying

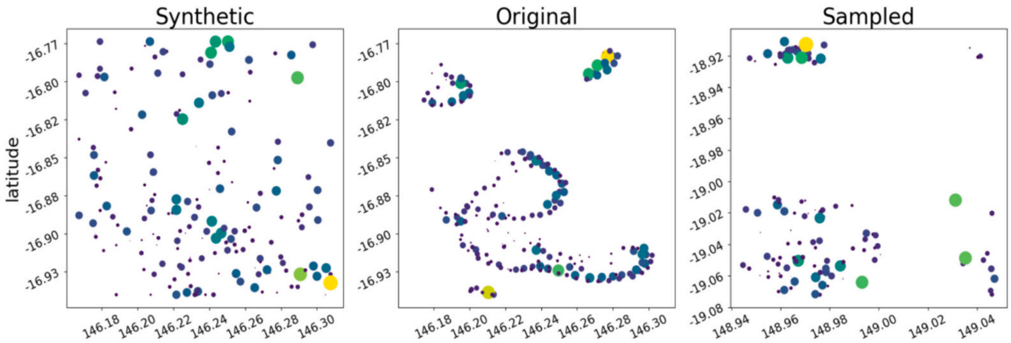


Fig. 3. Spatial distributions of the original (middle), synthetic (left) and sampled synthetic (right) data sets. Colours and size of points indicate site area. Note, sampled synthetic uses anonymized latitudes and longitudes.

real from fake data. This process is repeated until the discriminator is sufficiently unable to discriminate between real and synthetic data, which is determined by a minimum number of misclassifications predicted by the discriminator's classifier.

Through consultation with a connectivity specialist, two additional datasets were appended to the connectivity data for model learning, a north-south distance matrix and an east-west distance matrix. These distances (formulae available in Appendix C) give a measure of the distance between sites in the direction of north-south and east-west components of tides and currents, which should impact the strength of connectivity between sites. Latitudes and longitudes for the receiving sites were also added to the dataset to allow use of nearest neighbours to select synthetic connectivity data consistent with the synthetic site data. Nearest neighbours (using the Haversine distance) had to be used in the connectivity data as conditional sampling with the GAN model was restrictively time consuming. To improve the accuracy of the nearest neighbours approach, a large sample is drawn with many more sites than the final synthetic site data before the latitudes and longitudes from the synthetic site data are used to find nearest neighbours in this larger synthetic sample. The large sample size means the sample effectively covers the sample space and assures the synthetic data is estimated from values at very nearby sites.

6. Case study: Synthetic data for the Moore Reef Cluster

A synthetic data package based on the Moore reef cluster is used to demonstrate the utility of the synthetic data models in emulating the statistics of the original data set. Note that in the figures comparing "Original", "Synthetic" and "Sampled" data, the "Sampled" data set is a synthetic data set which has been conditionally sampled to be at the latitudes and longitudes of the sites generated in the synthetic site data, while "Synthetic" is generated at randomized latitudes and longitudes in the domain.

6.1. Site data

A sample of 108 sites was generated from the site data model. First a set of 200 sites was drawn from the unconditional model, then a set of site latitudinal and longitudinal centroids were sampled in randomized radii around 10 of the sites in the 200-site set. Finally, the conditionalized model is used to simulate the site data for the final 108 site set. The positions of the original, synthetic, and synthetic conditionally sampled sites are shown in Fig. 3, with the colour and size of the points indicating relative site area. Comparing the three figures, the model does not effectively replicate the spatial clustering of sites found in the original dataset. Clustering is better represented in the final sampled dataset, where positions are chosen in a randomized radius around a subset of the original sites, although still not a perfect representation of site spatial distributions.

The distributions of key variables in the original, synthetic, and sampled datasets are compared in Fig. 4, including the site area (m^2), percentage of the site area available for coral cover (k) and the median site depth (m). The synthetic site data achieves a mean quality score of 90%, as summarised in Table 2. The breakdown of this quality score in Table 2 shows that the model best captures the distributions of area, k, median depth, latitude, and longitude. From Fig. 4, the synthetic data generally captures the bi-modality of variables such as k and mean depth, although sites with very high k are underrepresented in the synthetic data, while sites with medium k are slightly over represented. Table 2 also details the diagnostics scores for the synthetic site data. The high synthesis and boundaries scores show that the data captures the range of values present in the original dataset well while being adequately synthetic (not being an exact copy of the original data). The quality and diagnostics scores suggest the synthetic data adequately

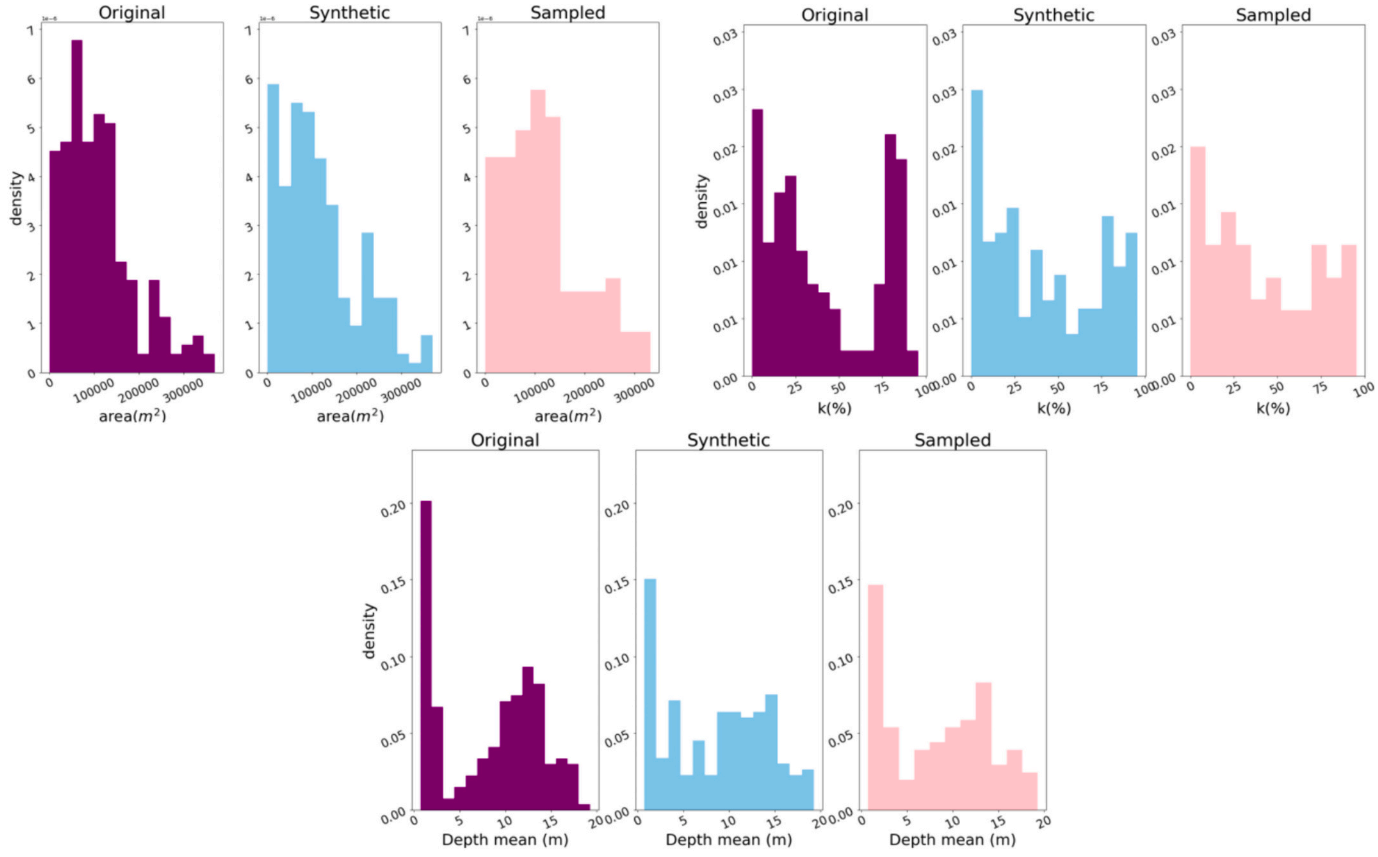


Fig. 4. Histograms comparing key variables in the site data set for the original, synthetic and sampled synthetic site data sets.

represents the statistics of the original dataset and would demonstrate a desirable range of environmental conditions for the purpose of model and method demonstrations and testing.

6.2. Cover data

A breakdown of the distributions of the cover data by species for the original, synthetic, and synthetic sampled datasets is shown in Fig. 5. Generally, the synthetic data emulates the species proportions effectively. The model does tend to overestimate species 4 and underestimate species 2, although this could be due to spatial distributions of the species coming into play. Histograms of the distribution of cover are shown in Fig. 6, suggesting the range of cover magnitudes and shape of the distribution is well captured, although covers above 0.03 are slightly underrepresented. The synthetic cover data gives a mean quality score of 90% in Table 2, with key variables cover and species scoring in the 89–90% range. The cover data also performs well in the diagnostics scores, rating highly on both coverage and boundaries in Table 2. The score for synthesis is lower, but still suggests the data is adequately synthetic for testing and method demonstration purposes.

6.3. DHW and Wave data

The original and synthetic DHW trajectories over time are compared in Fig. 7. On inspection, although the distribution of DHW values is a lot smoother for the sampled synthetic data, the range and spread of values is reasonably well captured, with the median DHW growing from around 2.5 at 2025 to around 15 in 2099. The maximum values are also similar, ranging from around 7 in 2025 to around 22 in 2099. The DHW data gives a mean quality score of 93%, with scores greater than 90% for DHW, latitude and longitude and greater than 85% for the column pairs DHW and year, DHW and latitude and DHW and longitude. The DHW data also performs generally well in the diagnostics scores, with high coverage and boundaries scores. The synthesis value is lower than the other datasets, suggesting high similarity to the original dataset. This is not so much an issue for the DHW data, as, unlike having a highly similar site data set, with anonymised geographical locations the DHW data should not be easily connected to the original despite high similarity.

The original and synthetic wave trajectories over time are compared in Fig. 8. Here again, the sampled and synthetic wave trajectories are a lot smoother than the original data set and the median is slightly lower for the synthetic datasets in later years than the original, although this could be a spatial effect. The range of the synthetic and sampled data sets also differ, with the synthetic data hitting the maximum value of 2 m with lower frequency. The data quality mean score is 87%, which is mostly reduced by the shape of the *Ub* (significant wave height) distribution poorly capturing the maximum values of the original data set. The data quality is fine for the current purpose of model testing and demonstration purposes, but the synthetic model could be improved by further refinement of training parameters. The *Ub* data performs well in the diagnostics scores, achieving high coverage, boundaries, and synthesis scores. The high boundaries score seems to contradict the plot of the synthetic *Ub* data, but this is because although the *Ub* data covers the range of values in the original data, the maximum values of the original data do not occur as frequently in the synthetic data. This could be due the synthetic data having half the number of sites as the original and also a different spatial configuration.

6.4. Connectivity data

The original and synthetic connectivity data are more difficult to compare due to being a relational data set. The quality score in Table 2 is listed for a comparison of an un-conditionalised sample of the connectivity data (sampled at the original site positions), because the columns of the original and synthetic data set must match for this metric. Due to this the quality ratings are particularly high for the connectivity data set, which would normally not be desirable as it would suggest a lack of model generalization. The sampled synthetic connectivity data (in the new synthetic site positions) is also evaluated using a visual comparison of pairwise correlations, in Fig. 9, and a visual comparison of correlation matrices in Fig. 10.

From Table 2, the synthetic data sampled at the original data site positions performs very well on the quality scores, both in terms of correlation relationships and capturing the distributions of individual sites connectivity with other sites. From the plotted PCA components in Fig. 9, the synthetic data components look like a slightly rotated version of the original data's components, which can be attributed to the first feature having greater weight in synthetic data, while the first three components are more equally weighted in the original data. Fig. 10 compares pairwise correlations across columns for the original, synthetic, and sampled connectivity data, showing that correlation relationships are well captured for both the synthetic un-conditionalized and synthetic conditionalized data set. The model overall captures the correlations and distribution of the data well while having some bias towards primary components of the data in the principal component analysis.

6.5. Case study outputs

Although the main purpose of the synthetic data developed here is to

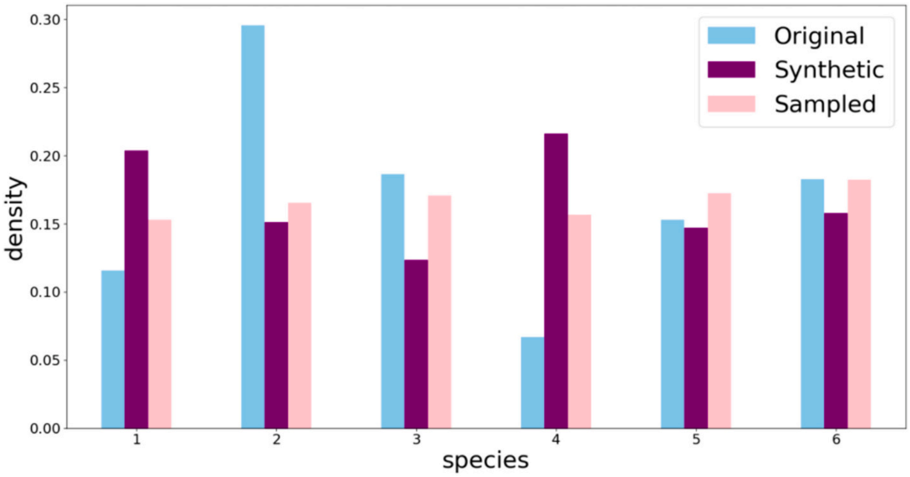


Fig. 5. Stacked bar cart comparing the proportional coral cover for each of the 6 species summed over size classes in the original, synthetic and synthetic sampled datasets.

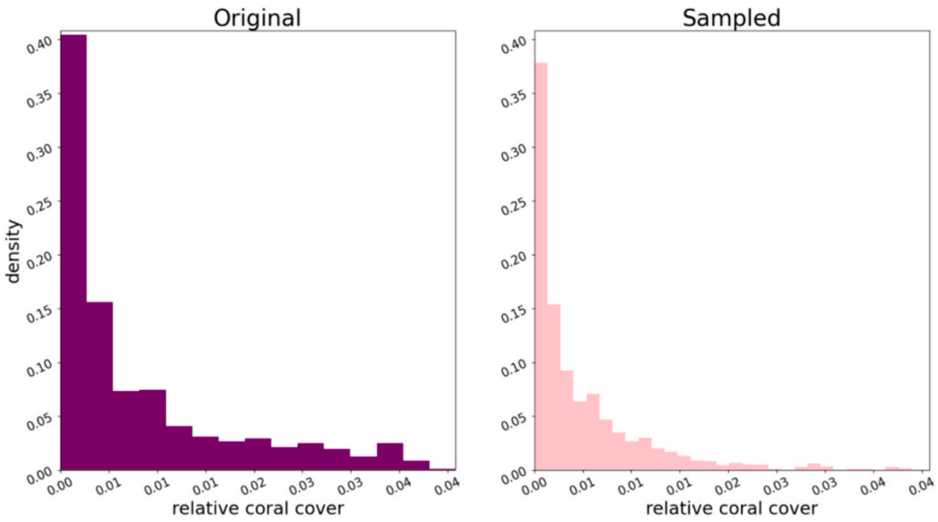


Fig. 6. Histograms comparing the distribution of coral cover across sites and species.

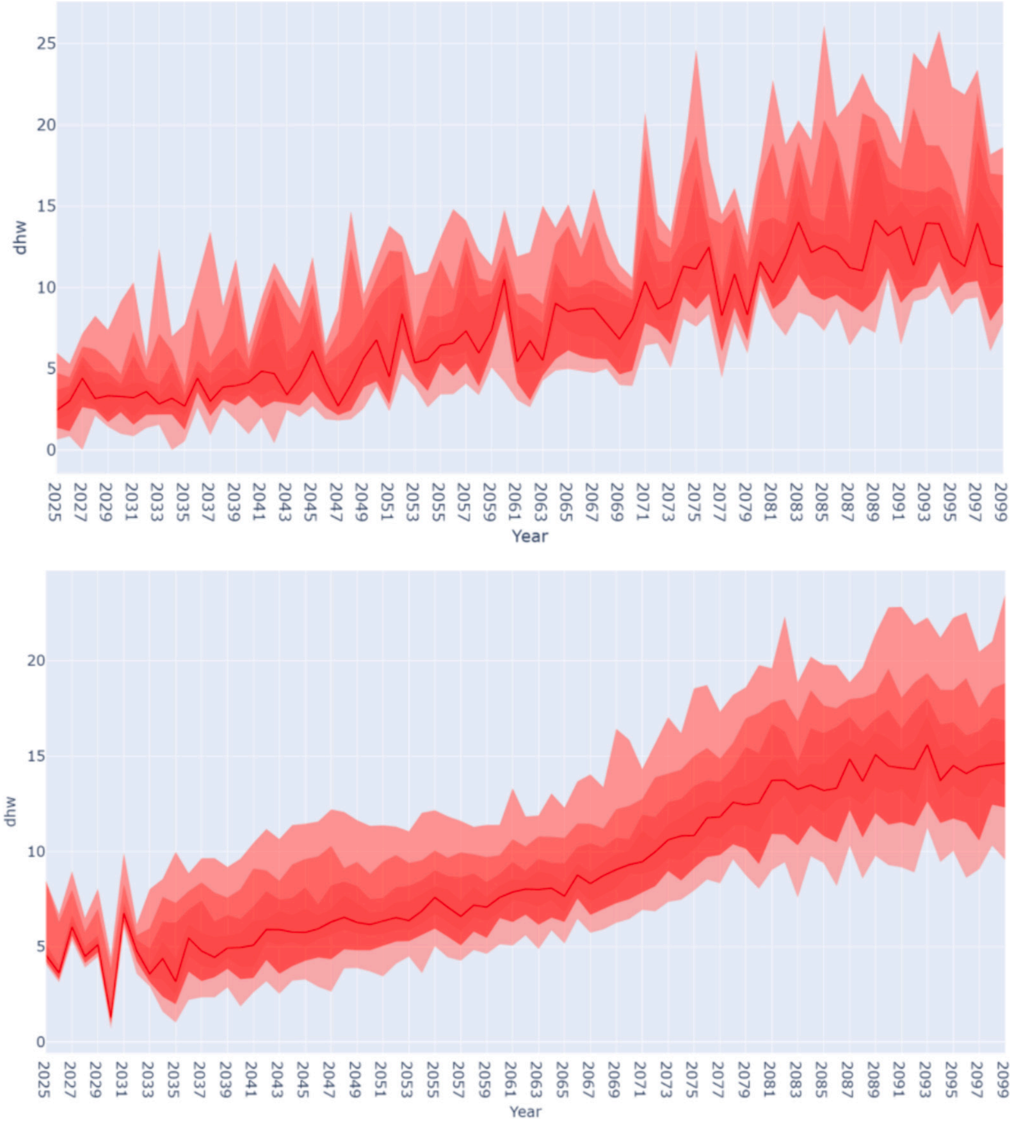


Fig. 7. Timeseries plots of the original (top) and synthetic sampled (bottom) DHW datasets. Red lines are the median across sites, with shading indicating the 25th, 50th and 75th quantiles.

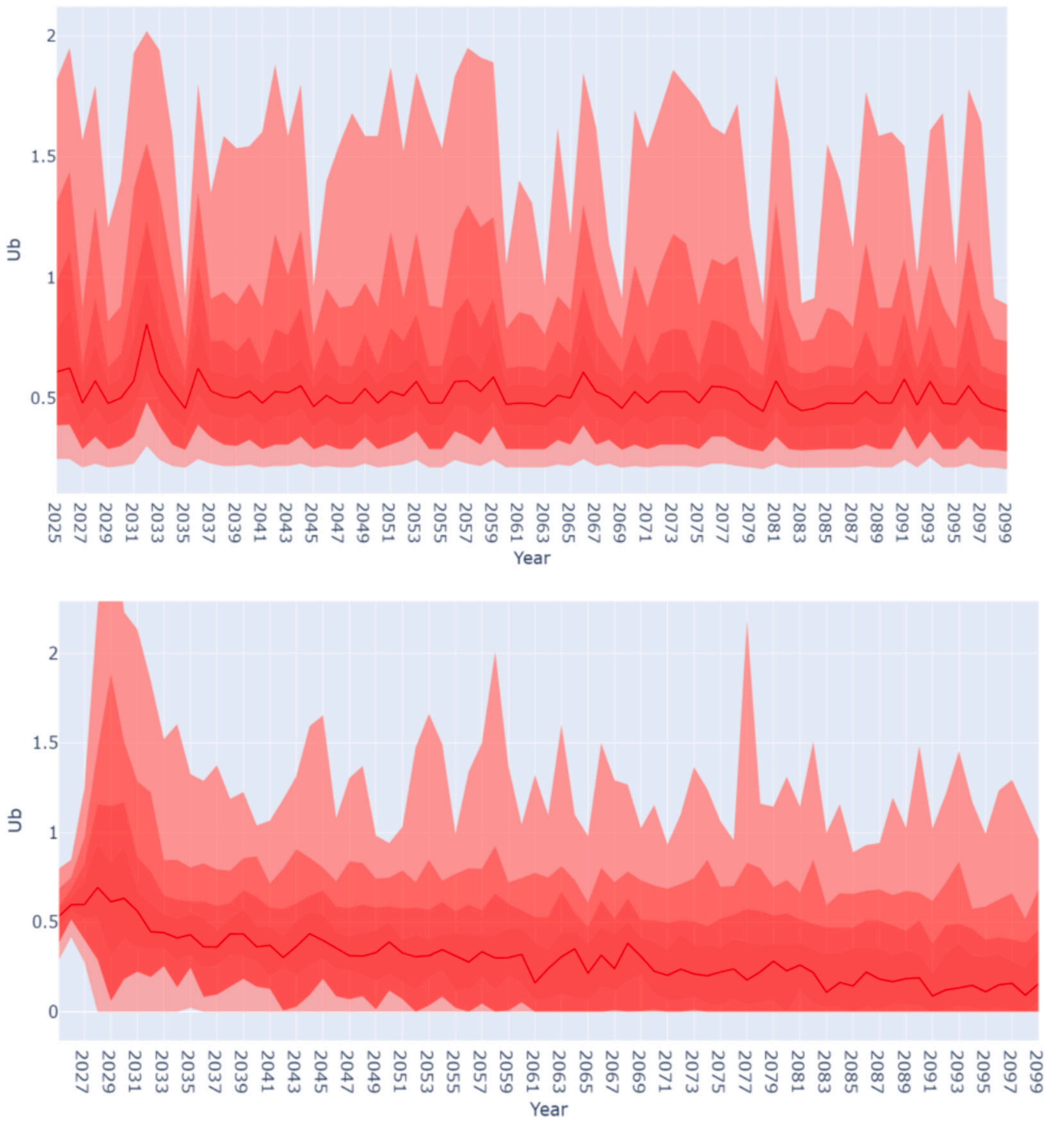


Fig. 8. Timeseries plots of the original (top) and synthetic sampled (bottom) wave height (Ub) datasets. Red lines are the median across sites, with shading indicating the 25th, 50th and 75th quantiles. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

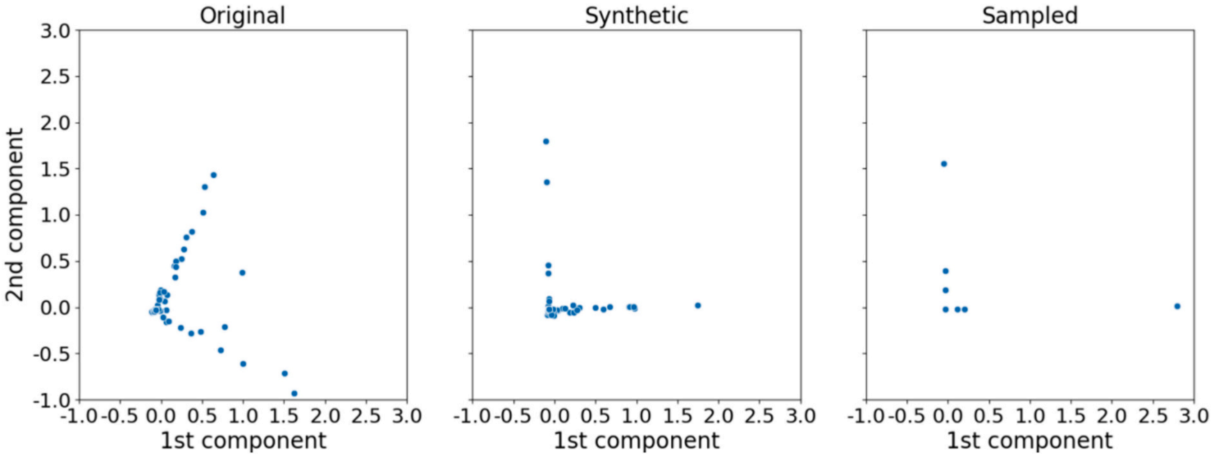


Fig. 9. Plots of first 2 PCA components for the original, synthetic and synthetic sampled datasets.

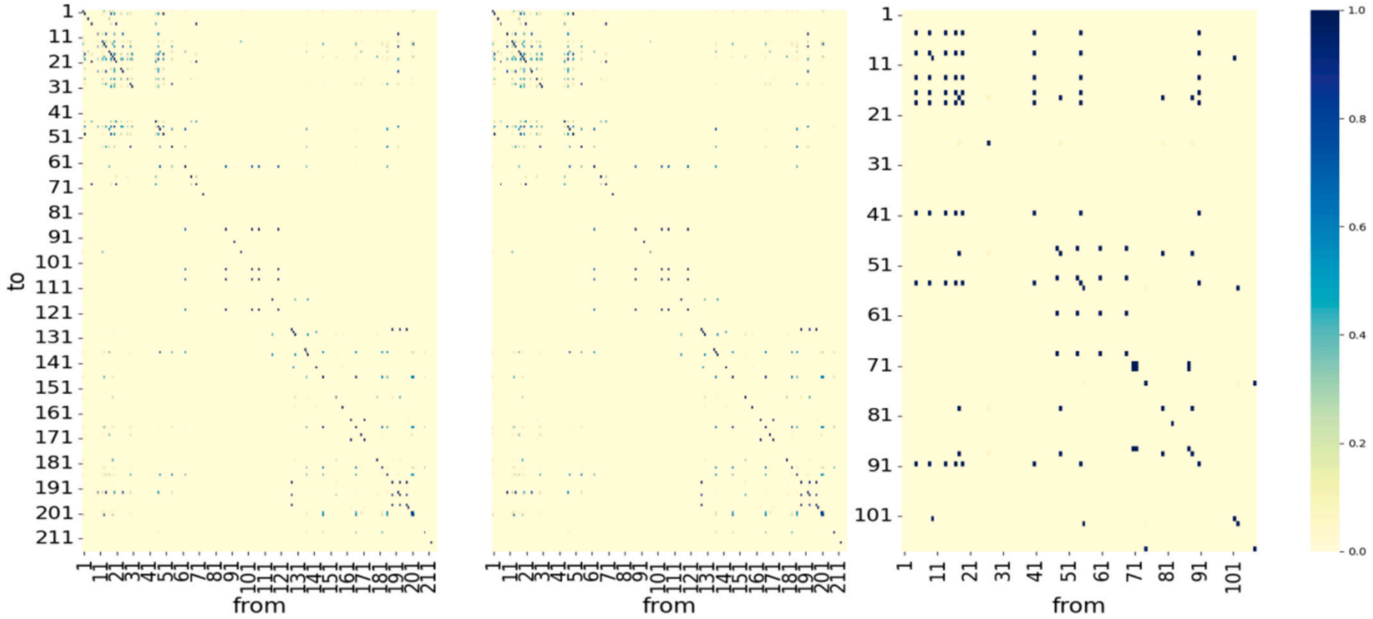


Fig. 10. Pairwise correlation matrices for the original (left), synthetic (middle) and synthetic sampled (right) datasets. Note the right most plot appears sparser because the synthetic data set contains half as many sites as the original.

test, develop and analyse ADRIA's site selection algorithms, we would expect the unaugmented synthetic data to display a similar magnitude and range of coral cover outcomes to the original data when run over many stochastic scenarios (with key ecological and environmental parameters sampled from their distributions). Fig. 11 shows comparisons of the total absolute coral cover over time output from the ADRIA model for 4096 scenarios using the original Moore domain and the synthetic datasets. As can be seen, the synthetic data captures the range and spread of outcome trajectories generally well. The synthetic data does demonstrate less variation, but this is likely because the synthetic data was based on 10 samples of 5 DHW replicates, rather than the full 50 replicates, which could reduce observed variability. The original data also has a higher initial coral cover, but this is due to the smaller number of sites in the synthetic data set.

6.6. Adding relationships between datasets

The datasets simulated here may have relationships between them which are difficult to capture when simulating each dataset with a different model. The challenge of simulating all the datasets together is a difficult statistical problem to address as the potential relationships between datasets are largely unknown and may be highly complex and non-linear. Fully investigating these relationships should be the focus of another study, however, some relationships may be easily integrated into the models as they are. Some relationships which are likely important include that between connectivity and coral cover, and between DHWs and cover. Here, we add these variables as predictors in the initial coral cover model and examine the results for improvement.

Connectivity, as summed incoming connectivity, and the mean DHW at each site at the start year were appended to the coral cover data during model training. These variables from the synthetic connectivity and DHW data were then used to conditionally sample the synthetic coral cover data, along with the synthetic latitudes and longitudes. This gave an overall quality score of 85.26% with column shapes score of 86.89% and column pair trends score of 83.63%, a reduced columns shapes score, and column pair trends score relative to training the model without these additional variables. On assessing the breakdown of the column pair trends score, the new model does capture correlations between connectivity and cover well (the correlation similarity score is 0.997), but captures the shape of the connectivity distribution poorly,

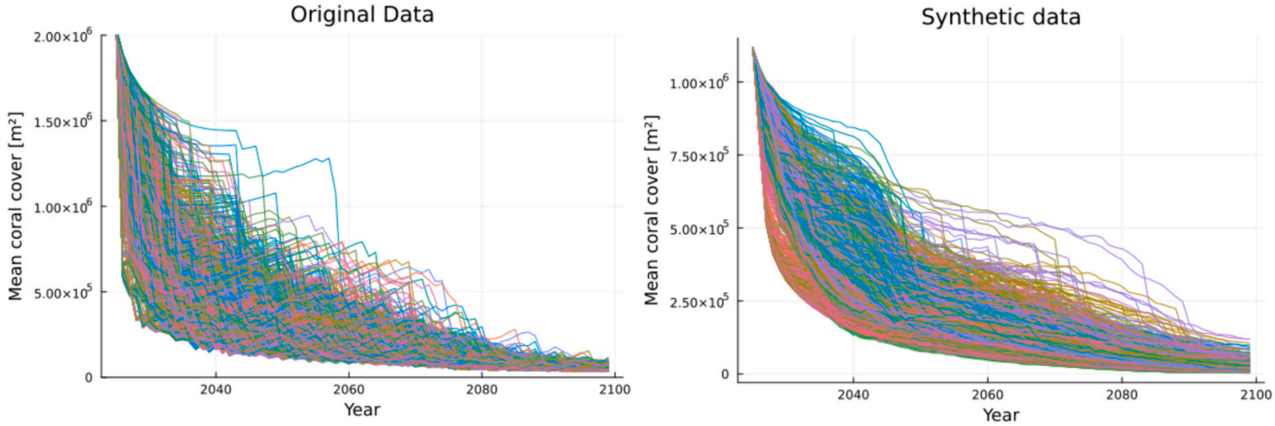


Fig. 11. Plots of coral cover (m^2) over time for 4096 randomized scenarios in ADRIA, using the original data inputs (left) and synthetic (right). Note the difference in magnitude is due to the synthetic data set having half as many sites as the original.

with an individual column shape score of 0.771. The correlations between DHWs and cover are also well captured (the correlation similarity score is 0.968), as well as the distribution shape of DHWs (the individual variable score is 0.961). However, the individual column shape score for cover is reduced (93.2% vs 88.0%), suggesting the data selected are not ideal for representing spatial relationships with coral cover. The model also took 2.5 h to run, which is 16 times that of the first model, due to the computational expense of conditionally sampling an additional two variables over ~100 sites.

This does not necessarily suggest that adding additional variables to inform spatial relationships with coral cover is unhelpful but that additional investigations into which relationships should be captured is needed. For example, we did not have access to hindcasted DHWs for this investigation and so instead used those for the first year of simulation, but past DHWs would probably best explain the spatial variations of the coral cover on the first year of simulation. Similarly, connectivity data is highly variable (both on a year to year and day to day scale), and so it is likely more informative to use some time integrated measure of strongest connections and variability of connections to capture relationships with spatial variations in cover. In both cases implementing improved representations of DHWs and connectivity for model training will take further investigation and may form further studies. For the purpose of testing and validation of the site selection algorithms, the models developed serve their purpose, but for other uses of the synthetic data better representation of spatial relationships between variables in different data sets is being investigated for future studies.

7. Discussion/conclusions

The synthetic data models developed here create a pipeline for generating synthetic reef data for the ADRIA decision support tool, suggesting approaches to create fit-for-purpose synthetic data pipelines for reef decision support applications. The data generally performs well in the case study against a variety of diagnostics, while also suggesting improvements for the synthetic data models.

The spatial distribution of the synthetic sites could be improved in several ways, such as by allowing sampling of sites across randomized arcs in the spatial domain, rather than circles. General correlations between benthic properties such as depth and features such as “reef slope”, “reef flat” etc. are learnt by the model for the synthetic site data, although perhaps a separate model should be trained for learning relationships between the spatial clustering of reefs and benthic information, which would allow spatial clustering patterns to be generated based on these properties. For the current purpose of creating test data for and demonstrating the ADRIA decision support tool, however, this additional complexity is not necessary.

The PAR model for wave data also does not capture the frequency of extremes of the wave height dataset well, suggesting that the model simulation and sampling parameters may need to be further optimized to generate better synthetic wave data. This could be done by testing different pre-learning data transformations and refining the model constraints based on the original data maximum and minimum values. The wave data quality is fine for the purpose of site selection testing and demonstrations but would need to be improved if it was to be used for a purpose which required higher fidelity to the physical dynamics, such as for sensitivity analysis.

As a first conceptual framework for a synthetic reef data pipeline the models developed here serve their purpose in facilitating decision support tool testing and demonstrations during development, although alternative models could be investigated to reduce the time taken to create some of the synthetic datasets and improve fidelity to the ecological dynamics if needed for other applications. Machine learning and neural network-based models are ideal for fitting models to a wide range of data sets but can take time for model training. The models for site data, coral cover and connectivity all show reasonable runtimes, but the PAR environmental data model’s runtime could be improved upon (it takes around 2.3 mins for one climate replicate for 200 sites). This is largely because a new model is trained for each climate replicate of the data and so with 20 climate replicates, the runtimes can be up to an hour. This is better than the original model the data comes from, which can take days, but would be desirable to improve if possible. Several alternative synthetic models were investigated during the development of the synthetic data repository for ADRIA, including y-data-synthetic’s TimeGAN and DoppleGANger models (YData, 2023), and pyunicorn’s time series surrogate models (Donges et al., 2015), however, the PAR model was settled upon as it was the only timeseries synthetic data model reviewed which offered conditional sampling. Although these alternatives don’t offer conditional sampling, interpolation methods could possibly be used in place of conditional spatial sampling if the outputs perform sufficiently well in replicating the statistics of the original data and offer significant runtime reductions.

A pertinent question which arises is “does the full set of synthetic input data sets successfully represent the full set of original data sets?”. This is a gnarly statistical problem which is usually approached in the context of synthetic data for machine learning problems via a prediction success rate (e.g. (Boyeau et al., 2024)). This this context, a model is trained on the synthetic data and the success of the model in predicting what it is designed to predict is evaluated. The same is repeated for the real data and prediction success rates are compared. In the case of our decision support tool, however, small differences in initial coral cover, DHW data and connectivity can have a significant impact on the ecological model’s output trajectory, so even for a set of data which may be statistically similar to the original data set, the outcomes may be quite different. We would, however, expect a similar spread and shape in the temporal trajectory of absolute coral cover, which we have compared here in the case study in section VI.

It is important to reiterate here that the synthetic data is not intended to be used for policy purposes or particular instances of reef management decision making. The original data will always be used for decision making as when a policy has been decided upon the real data will be published according to official release processes. Also, management decisions are highly dependent on spatial context and it would be inappropriate to use synthetic data in this capacity. The synthetic data pipeline developed here is intended for testing, validating and demonstrating ADRIA’s decision support tool functionality, particularly in the case of supporting publications demonstrating the model’s decision support capability without prematurely revealing results of an actual reef dataset which could have implications for partners, rightsholders and stakeholders. In line with this purpose, sufficient emulation to demonstrate and test the ADRIA decision support tool is the goal, not perfect emulation of the original dataset down to the exact ecological, biophysical, and dynamical conditions. The data would not yet, however, be suitable for use in sensitivity analyses of model outcomes to input data layers, which is an eventual goal of the synthetic data. For the data to be suitable for such an application, further investigations into relationships between datasets, such as the impact of connectivity on coral cover and vice versa should be investigated and attempted to be represented in the synthetic data generators. Due to the complexity and non-linearity of such relationships this would require its own in-depth study involving connectivity and DHW specialists and, although will pursued in further publications, is outside the scope of this paper.

Despite the drawbacks discussed, a data pipeline for producing on-demand synthetic data packages for methods testing, validation and publishing demonstrations for the reef decision support tool ADRIA is developed. Using the accessible and flexible synthetic data libraries SDV and y-data-synthetic, we demonstrate that synthetic data for a reef decision support tool can be developed for decision support tool datasets without requiring building models from the ground up for each dataset. With the push for embracing big data in ecological and environmental modelling, synthetic data for reef decision support tools offers a means of fully capitalising on big data models and building a rigorous testing, development, and publication pipeline where data is scarce or sensitive.

CRediT authorship contribution statement

Rose Crocker: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Investigation, Formal analysis, Data curation, Conceptualization. **Barbara J. Robson:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Chinenye Ani:** Methodology, Conceptualization. **Ken Anthony:** Writing – review & editing, Supervision. **Takuya Iwanaga:** Writing – review & editing, Methodology, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The synthetic data can be made available but the original data cannot. All code has been made available and is cited in the paper.

Appendix A. Descriptions of original dataset creation

A.1. Connectivity data

Larval connectivity data represents how coral larvae travel between different locations on a reef due to ocean currents and other factors. Connectivity is determined by a variety of factors, including ocean currents, larval development rates, mortality, and settlement. The larval connectivity data used here is generated using a particle tracking simulator, called OceanParcels (Van Sebille et al., 2020) with outputs from a hydrodynamic marine model implemented using RECOM (The Great Barrier Reef Authority, 2022) within a lower (1 km) resolution hydro-dynamic model of the GBR (version 2.0 of the eReefs hydrodynamic model for the GBR (Steven et al., 2019), an implementation of SHOC (Herzfeld and Rizwi, 2019). Nesting the RECOM within a lower resolution model allows particles which travel outside the higher resolution model's boundaries and then return to still be tracked. Details of the creation of ADRIA's connectivity dataset's can be found in (Ani et al., 2023).

A.2. DHW data

Degree Heating Weeks (DHWs) are a measure of the cumulative heat stress experienced by corals over a time period. Degree heating weeks are calculated by comparing temperatures predicted by RECOM with average water temperature for each location and summing the integrated deviation above the climatological value over a 12-week period. This running 12-week integral is calculated over the course of each October to February model run, and the maximum value is calculated from the resulting time series for each grid cell. Two climatology data sources allow for calculation of two alternative DHW products from the same RECOM run. These are the methods developed and tested for the GBR by Langlais et al. (2021) and by Wijffels et al. (2018). Only surface DHW is used by the ADRIA model, and coral bleaching responses are adjusted as a function of depth.

A.3. Wave data

Routine wave exposure is modelled by forcing a near-shore numerical wave model (UQ SWAN) with wind data through a high-resolution representation of bathymetry to define the three-dimensional underwater reef structures and surrounding depth profile. Maps of exceedance probabilities are then generated from the model runs to show the wave heights that would be expected not to be exceeded in each of several percentages of the time. ADRIA's wave height data uses the 90th percentile (Callaghan et al., 2015).

Appendix B. Data variable descriptions

Dataset	Variable descriptions
Site data	Site id: A unique identifier for a particular area of reef. Each reef or group of reefs is split up into polygons and each polygon is given an ID. K: The percentage of the total site area which can foster coral growth (also called the carrying capacity). Area: The area of the site polygon in m^2. Zone: The GBRMPA zone to which the site belongs (Great Barrier Reef Marine Park Authority, 2022) Coral algae: Proportion of the site covered with coral algae. Rock: Proportion of the site covered with rock. Habitat: 'Crest', 'Outer Flat', 'Sheltered Slope' or 'Slope', describing the site habitat. Rubble: Proportion of the site covered with rubble. Sand: Proportion of the site covered with sand. Depth median: Median depth of sites in m. Depth mean: Mean depth of sites in m. Depth standard deviation: Standard deviation of depth across each site. Site polygons: Describe the geometry of each site as polygon of latitudinal and longitudinal coordinates. The centroids of these polygons are used to train the synthetic site data model.
Initial coral cover	Site: Integer corresponding to a site id. Species: Integer from 1 to 36 designating the size and species of coral. Coral cover: Proportion of coral cover on each site at the initial time step of the model.
DHW data	Site: Integer corresponding to a site id. Year: Datetime from 2025 to 2099. DHW: Degree heating weeks for each year and site. One DHW is equivalent to one week of sea surface temperature 1 degree Celsius above the expected climatological value (the long-term mean of the maximum value for that month of the year). One week at 2 degrees above the climatological value would accrue 2 DHWs. DHWs are cumulative over a 12-week summer period for each year (National Oceanic and Atmospheric Association, 2023)
Wave data	Site: Integer corresponding to a site id. Year: Datetime from 2025 to 2099. *Ub*: 90th percentile significant wave heights (the height expected not to be exceeded 90% of the time).

(*continued on next page*)

(*continued*)

Dataset	Variable descriptions
Connectivity data	Connectivity: The biophysical connection between a set of sites. In the context of coral larval connectivity, this is expressed as a sparse matrix representing the probability of larvae released a from each site reaching each other site.

Appendix C. Tidal distance formula

North-south and East-west tidal distances were appended to the connectivity data when training the synthetic connectivity data model as it was found to perform better in terms of capturing spatial relationships in the connectivity data.

The formula used for tidal distance between each pair of sites is:

$$d = R\left(2arctan\left(\sqrt{a},\ \sqrt{1-a}\right)\ \right),$$

where

$$a = \left(sin\frac{b_1 - b_2}{2}\right)^2,$$

and

$$b_1 = \pi l_1/180, b_2 = \pi l_2/180.$$

R is the radius of the Earth. For the East-West tidal distance, l_1 and l_2 are the longitudes of the pair of sites, while for the North-South tidal distance l_1 and l_2 are their latitudes. This gives two matrices of size (number of sites) by (number of sites).

Appendix D. Model comparisons

Data	Model Type	Quality scores	Total time (s)	Time per site (s)
Site data	Gaussian Copula with default distribution = Normal	Overall Quality Score: 85.56% Column Shapes: 83.15% Column Pair Trends: 87.98%	53.06	0.27
	Gaussian Copula with default distribution = Gaussian KDE	Column Shapes: 88.81% Column Pair Trends: 90.97% Overall Quality Score: 89.74%	68.77	0.34
	TVAE	Overall Quality Score: 85.13% Column Shapes: 82.33% Column Pair Trends: 87.93%	85.76	0.43
	CTGAN	Overall Quality Score: 79.33% Column Shapes: 79.94% Column Pair Trends: 78.72%	102.6	0.51
	CopulaGAN	Overall Quality Score: 72.86% Column Shapes: 69.34% Column Pair Trends: 76.39%	110.93	0.55
Coral cover data	Gaussian Copula with default distribution = Normal	Overall Quality Score: 86.71% Column Shapes: 86.91% Column Pair Trends: 86.51%	69.98	0.35
	Gaussian Copula with default distribution = Gaussian KDE	Overall Quality Score: 90.84% Column Shapes: 91.9% Column Pair Trends: 89.78%	88.60	0.44
	TVAE	Overall Quality Score: 90.7% Column Shapes: 92.8% Column Pair Trends: 88.59%	492.31	2.46
	CTGAN	Overall Quality Score: 82.55% Column Shapes: 83.23% Column Pair Trends: 81.86%	523.07	2.61
	CopulaGAN	Overall Quality Score: 79.11% Column Shapes: 79.19% Column Pair Trends: 79.03%	614.19	3.07
DHW data	PAR Model	*Column Shapes:* 95.34% *Column Pair Trends:* 89.97% *Overall* Quality *Score:* 92.66%	145.28 (for each climate replicate)	0.72
Wave data	PAR Model	*Column Shapes:* 87.34% *Column Pair Trends:* 86.71% *Overall Quality Score:* 87.03%	145.62 (for each climate replicate)	0.73
Connectivity data	GAN	Column Shapes: 98.2% Column Pair Trends: 99.9% Overall Quality Score: 99.0%	165.37	0.83
	Gaussian Copula with default distribution = Normal	Overall Quality Score: 91.2% Column Shapes: 85.25% Column Pair Trends: 97.14%	22.19	0.11

(*continued on next page*)

(*continued*)

Data	Model Type	Quality scores	Total time (s)	Time per site (s)
	Gaussian Copula with default distribution = Gaussian KDE	Overall Quality Score: 91.82% Column Shapes: 86.25% Column Pair Trends: 97.39%	48.94	0.24

References

Alizadeh, R., Allen, J., Mistree, F., 2020. Managing computational complexity using surrogate models: a critical review. Res. Eng. Des. 31, 275–298. https://doi.org/10.1007/s00163-020-00336-7.

Allken, V., Handegard, N., Rosen, S., Schreyeck, T., Mahiout, T., Malde, K., 2019. Fish species identification using a convolutional neural network trained on synthetic data. ICES J. Mar. Sci. 76 (1), 342–349. https://doi.org/10.1093/icesjms/fsy147.

Ani, C. J., Haller-Bull, V., Gilmour, J., & Robson, B. (2023). Connectivity modelling at local scales identifies sources and sinks of coral recruitment within reef clusters. Submitted to journal X.

Berry, K.J., Johnston, J.E., Mielke Jr., P.W., Johnston, L.A., 2018. Permutation Methods. Part II. Comput. Stat. 10 (3), e1429 https://doi.org/10.1002/wics.1429.

Boyeau, P., Angelopoulos, A.N., Yosef, N., Malik, J., Jordan, M.I., 2024. AutoEval done right: using synthetic data for model Evaluation arXiv, pp. 1–12. Retrieved from arXiv:2403.07008v1.

Callaghan, D.P., Leon, J.X., Saunders, M.I., 2015. Wave modelling as a proxy for seagrass ecological modelling: comparing fetch and process-based predictions for a bay and reef lagoon. Estuar. Coast. Shelf Sci. 153, 108–120.

Chen, R.J., Lu, M.Y., Chen, T.Y., Williamson, D.F., Mahmood, F., 2021. Synthetic data in machine learning for medicine and healthcare. Nat. Biomed. Eng. 5, 493–497. https://doi.org/10.1038/s41551-021-00751-8.

Cresswell, A.K., Haller-Bull, V.C.-R., Gilmour, G., Bozec, Y.-M., Barneche, D., Robson, B., Ortiz, J.-C., 2023. Modelling coral population dynamics at within reef scales in a mechanistic framework. Draft 1–37.

DataCebo, I., 2023, April. Synthetic Data Metrics. Version 0.9.3. Retrieved from https://docs.sdv.dev/sdmetrics/.

Donges, J., Heitzig, J., Beronov, B., Wiedermann, M., Runge, J., Feng, Q.-Y., Kurths, J., 2015. Unified functional network and nonlinear time series analysis for complex systems science: The pyunicorn package. Chaos 25 (113), 101. https://doi.org/10.1063/1.4934554.

Endres, M., Venugopal, A., Tran, T., 2022. Synthetic data generation: a comparative study. In: 26th International Database Engineered Applications Symposium. Association for Computing Machinery, New York, pp. 94–102. https://doi.org/10.1145/35487885.3548793.

Ernst, M., 2004. Permutation methods: a basis for exact inference. Stat. Sci. 19 (4), 676–685. https://doi.org/10.1214/088342304000000396.

Fassnacht, F.E., Latifi, H., Hartig, F., 2018. Using synthetic data to evaluate the benefits of large field plots for forest biomass estimation with LiDAR. Remote Sens. Environ. 213, 115–128. https://doi.org/10.1016/j.rse.2018.05.007.

Great Barrier Reef Marine Park Authority, 2022, August 23. Zoning Maps. Retrieved from. https://www2.gbrmpa.gov.au/access/zoning/zoning-maps.

Herzfeld, M., Rizwi, F., 2019. A two-way nesting framework for ocean models. Environ. Model. Softw. 117, 200–213.

Iwanaga, T., Anthony, K.R., Robson, B., Crocker, R., Ani, C., Ribeiro De Almeida, P., Michael, T., 2024. Adaptive Dynamic Reef Intervention Algorithm. https://doi.org/10.5281/zenodo.7879777.

James, S., Habron, C., Branson, J., Sundler, M., 2021. Synthetic data use: exploring use cases to optimise data utility. Discover Artif. Intell. 1–15 https://doi.org/10.1007/s44163-021-00016-y.

Langlais, C.E., Herzfeld, M., Klein, E., Cantin, N., Benthuysen, J., Steinberg, C., 2021. Oceanographic drivers of bleaching in the GBR: from observations to prediction, Volume 2: 3D- Bleaching in the GBR: Development and analysis of a 3D climatology and 3D heat accumulation bleaching products using eReefs. Cairns: Report to the National Environmental Science Program. Reef and Rainforest Research Centre Limited.

Little, R.J.A., Liu, F., Raghunathan, T., 2004. Applied modelling and causal inference from incomplete-data perspectives. In: Little, R.J.A., Liu, F., Raghunathan, T., Gelman, A., Meng, X. (Eds.), Statistical Disclosure Techniques Based on Multiple Imputation. Wiley, New York, pp. 141–152.

Mahmood, A., Bennamoun, M., An, S., Sohel, F., Boussaid, F., Hovey, R., Kendrick, G., 2020. Automatic detection of Western rock lobster using synthetic data. ICES J. Mar. Sci. 77 (4), 1308–1317. https://doi.org/10.1093/icesjms/fsz223.

National Oceanic and Atmospheric Association, 2023. Satellites & Bleaching. Retrieved from NOAA satellite and information service. https://coralreefwatch.noaa.gov/product/5km/tutorial/crw10a_dhw_product.php.

Nikolenko, S.I., 2021. Synthetic Data for Deep Learning. eBook. Springer. https://doi.org/10.1007/978-3-030-75178-4.

Patki, N., Wedge, R., Veeramachaneni, K., 2016, October. The Synthetic Data Vault. In: IEEE International Conference on Data Science and Advanced Analytics (DSAA), pp. 399–410. https://doi.org/10.1109/DSAA.2016.49.

Poisot, T., Gravel, D., Leroux, S., Wood, S.A., Fortin, M.-J., Baiser, B., Stouffer, D.B., 2016. Synthetic datasets and community tools for the rapid testing of ecological hypotheses. Ecography 39, 402–408. https://doi.org/10.1111/ecog.01941.

Quintana, D., 2020. A synthetic dataset primer for the behavioural science to promote reproducability and hypothesis generation. Elife 9, e5327. https://doi.org/10.7554/eLife.5327.

Raghunathan, T., 2021. Synthetic data. Ann. Rev. Stat. Appl. 8 (1), 129–140. https://doi.org/10.1146/annurev-statistcis-040720-031848.

Reichman, O.J., Jones, M.B., Schildhauer, M.P., 2011. Challenges and opportunities of open data in ecology. Science 331 (6018), 703–705. https://doi.org/10.1126/science.1197962.

Snoke, J., Raab, G.M., Nowok, B., Dibben, C., Slavkovic, A., 2018. General and specific utility measures for synthetic data. J.R.Statist.Soc. Statistics in Society A 181 (3), 663–688. https://doi.org/10.48550/arXiv.1604.06651.

Steven, A., Baird, M., Brinkman, R., Car, N., Cox, S., Herzfeld, J., 2019. eReefs: An operational information system for managing the Great Barrier Reef. J. Operat. Oceanogr. 12 (2), S12–S28.

The Great Barrier Reef Authority, 2022. Retrieved from eReefs RECOM v2 Alpha. https://recom.ereefs.info/.

Van Sebille, E., Delandmeter, P., Lamge, M., Rath, W., Scutt Phillips, J., Kronberg, J., Sterl, M., 2020. OceanParcels/parcels: Parcels v2.0.0: A Lagrangian Ocean Analysis Tool for the Petascale AGe.

Watson, P., 2015. Development of a unique synthetic data set to improve sea-level research and understanding. Coastal Res. 31 (3), 758–770. https://doi.org/10.2112/JCOASTRES-D-14-00143.1.

Watts, J.D., Powell, S.L., Lawrence, R.L., Hilker, T., 2011. Improved classification of conservation tillage adoption using high temporal and synthetic satellite imagery. Remote Sens. Environ. 115 (1), 66–75. https://doi.org/10.1016/j.rse.2010.08.005.

Wijffels, S., Beggs, H., Griffin, C., Middleton, J., Cahill, M., King, E., Sutton, P., 2018. A fine spatial-scale sea surface temperature atlas of the Australian regional seas (SSTAARS): seasonal variability and trends around Australasia and New Zealand revisited. J. Mar. Syst. 187, 156–196.

Wilson, R., Speirs, D., Sabatino, A., Heath, M., 2018. A synthetic map of the north-west European Shelf sedimentary environment for applications in marine science. Earth Syst. Sci. Data 10 (1), 109–130. https://doi.org/10.5194/essd-10-109-2018.

Wimmer, S., Finger, R., 2023. A note on synthetic data for replication purposes in agricultural economics. J. Agric. Econ. 74, 316–323. https://doi.org/10.1111/1477-9552.12505.

Xu, L., Skoularidou, M., Cuesta-Infante, A., Veeramachaneni, K., 2019. Modelling Tabular Data using Conditional GAN. In: *NIPS'19: Proceedings of the 33rd International Conference on Neural Information Processing Systems, 659*, pp. 7335–7345.

YData, 2023, May. YData Synthetic. Retrieved from. https://pypi.org/project/ydata-synthetic/.

Zhang, K., 2021. On Mode Collapse in Generative Adversarial Networks. Artificial Neural Networks and Machine Learning – ICANN 2021. Springer Link, pp. 563–574.

Zhang, K., Pakti, N., Veeramachaneni, K., 2022. Sequential Models in the Synthetic Data Vault. ArXiv, 1–17. https://doi.org/10.48550/arXiv.2207.14406.