# GCMP Mapping File Column Definitions

In MixS format for the Earth Microbiome Project Submission

**Overall:** Our overall goal is to produce a mapping file that is machine readable in QIIME, and contains columns that would be useful for statistical testing or visualization. The requirements here are slightly more tricky than usual, because in addition to being valid QIIME files, the mapping should be formatted for upload to QIITA (the QIIME database), and include where possible fields from the Minimal Information about any (X) sequence standard (MIxS), which is specified by the EMP. Navigating all this may take a bit of back-and-forth with the EMP folks standardizing metadata, but if we can make a good effort I think it will help our case.

**Issues:** Places where I think we may need to revise to match external standards are noted in red.

**Suggestions for merging with Kiribati samples.** Are in green.

**Validation:** Mapping files can be validated using [validate\_mapping\_file.py](http://qiime.org/scripts/validate_mapping_file.html) in the QIIME package. This also helps correct some minor issues (see the output corrected text file).

**References:**

The QIIME mapping file format is a tab-delimited text file described [here](http://qiime.org/documentation/file_formats.html). Making sure our submission is a valid QIIME mapping file is the most important requirement.

The QIITA requirements for sample and prep templates are [here](http://qiita.microbio.me/static/doc/html/tutorials/prepare-templates.html#sample-template). Since the EMP is doing all of the sequencing, I believe that we want to include all the fields for the sample template, but not the prep template. In addition to the sample name field we want to include all the fields required for ‘EBI export’ and ‘Centralized QIITA’. These are included in the list of column definitions below.

## Notes

These may be obvious or too basic, so feel free to skip. Below I’ve kept a list of the common issues I’ve encountered translating dive notes or human-readable datasets into mapping files are noted below, with some potential solutions. Additional notes are in the column definitions.

**Spell out all species and genus names.** Human-readable data often uses rows above or below to fill in missing details, but scripts won’t be able to do that, so all rows must be self-contained and explicit.

**Fill blank cells.** For missing data we’ve been using ‘Unknown’ for measures that apply but weren’t collected for that sample, and ‘None’ for measures that weren’t relevant. Blank cells/columns also have insidious effects in Excel, since selections stop on blanks, potentially allowing a sort to scramble data relative to ids.

QIITA documentation recommends NA in place of None. Not sure this matters (their main point is not to leave it blank), but we can skip if needed.

**Decompose text descriptions where possible.** It’s often useful to keep the raw text, but it of course can’t be processed automatically. So we often break up descriptions that share common elements into categories. Sometimes quite a few columns are needed to capture relevant aspects of a text description. For example, narratives of algal contact get decomposed into a lot of binary columns about the presence or absence of algal contact, how many types of algae are touching corals. In the HERBVRE project, where algal contact was very important, we also had a ton of binary columns for the presence/absence of contact by common algae like *Dictyota*. This can quickly spiral out of control, so its a bit of a balancing act based on the study questions...

**Remove non-standard characters.** The mapping file definition sends some mixed messages about exactly which characters are currently acceptable, but in general its safest to stick with alphanumeric characters(a-z,A-Z,0-9), underscore (“\_”), period(“.”), and dash(“-”; for negative numbers). Tab characters inside cells are particularly fatal, as these will offset columns. **Obnoxiously, the #SampleID and sample\_name columns can’t accept underscores, so periods must be used there.** (I think this is because some scripts concatenate things onto the SampleID with underscores, and they need to be able to separate back out the old SampleID).

QIIME can distinguish between tabs and spaces, and increasingly can handle more unusual characters, but these often cause headaches in other software. I usually replace spaces with underscores in data entries to avoid these issues, although in some columns we’ve used spaces (where the output is likely to appear in plots) so this isn’t wholly consistent.

**Binary values as ‘y’ or ‘n’.** We’ve tried to consistently represent binary values as ‘y’ for Yes and ‘n’ for no (replacing ‘True’,’Yes’,’Present’,etc). Arguably could be more convenient for labels spelled out, or for import to some scripts as numeric 1 or 0, but this is compact, and it is easy to change later as long as we’re consistent.

## Column Definitions

**#SampleID.** This is, obnoxiously, the same as sample\_name, but for some reason QIITA wants sample\_name, whereas actually running scripts requires #SampleID. So we plan to just duplicate SampleID at the last step. **[See sample\_name below]**

**sample\_id** (for KI data)

Season prefix: KI14=August 2014, KI15a=January 2015, KI15b=May 2015, KI15c=July 2015

Sample Prefix: FSYM and FQ=coral fragments preserved in guanadinium buffer, FMD=coral fragments frozen, FF= coral fragments preserved in formaldehyde, WSYM=water samples preserved in guanadinium buffer, WMD=water samples preserved in RNA later, SSYM=sediment preserved in guanadinium buffer, SMD=frozen sediment

Example: KI15cFMD077: Kiritimati field season August 2014 fragment microbial DNA tube#077

**sample\_name.** This column is designed to meet with QIITA and MIxS requirements. Alphanumeric and periods only. For the Global Coral Microbiome Project, the sample ids are based on a combination of the expedition and coral sampled. Each column, listed below, is separated by a period.

Expedition number (E1 = expedition 1)

Coral clade (phylogenetic group based on the clade system of Fukami et al), as an integer.

Three letter genus name

Four letter species name

Replicate colony number (to distinguish colonies of the same species from the same expedition)

Concatenated data (yyyymmdd)

Additional notes on the GCMP SampleID Format Details:

*Coral samples*: expedition number, clade number (not in roman numerals), 3 letter genus name, 4 letter species name, individual number, concatenated date (yyyymmdd), then biological compartment (M = 'mucus', S='skeleton',T='tissue', W = 'whole' (for outgroups). *Outgroup samples*: expedition number, "Outgroup", full genus name, full species name or sp. or sp., sample\_number, concatenated\_date (yyyymmdd), biological compartment ((M = 'mucus', S='skeleton',T='tissue', W = 'whole').

*Water samples*: expedition number, "Water" sample number, concatenated date (yyyymmdd).

*Sediment samples*: expedition\_number, "Sediment", sample\_number, concatenated date (yyyymmdd)

**field\_sample\_name** As above, but without the suffix describing tissue separation. Once a coral or other benthic invertebrate is separated into tissue, mucus, and skeleton compartments, we add an M, S,T or W to each. This should be the same as sample\_name and #SampleID for water and sediment samples. **Note:** we’ve uploaded sample photos to Flickr with tags for each field\_sample\_id. So if you want to see id E4.17.Ech.lame.1.20150305

, you can search it on Flickr and get [here](https://www.flickr.com/search/?text=E1.3.Por.loba.1.20140724).

**expedition\_number** This is the number of the expedition for the Global Coral Microbiome Project. *Example:* E1, E8. So far the expeditions with numbers are: E1, Lizard Island 2014; E2, Bocas del Toro 2014; E3, Lizard Island 2015; E4, KAUST\_2015; E5, Cartagena, Colombia 2015; E6, Curacao; E7 Ningaloo, 2015; E8, Singapore 2015; E9, Reunion 2015; E10, Lord Howe 2015; E11, Pacific Panama. The following don’t have codes yet (mostly planned expeditions that haven’t happened so far): Brazil, Moorea, Indonesia, Panama\_Coiba\_Pearl\_Islands.

**sampling\_expedition** The short name of the sampling expedition. *Examples:* KAUST\_2015, Bocas\_del\_Toro\_2014.

**colony\_sample\_name** This isn’t very formally defined, but each expedition had some sort of system for generating unique ids for each colony. These might be on freezer labels etc, so I didn’t want to lose them, and stuck them in here. Should map back to the coral colony sampled uniquely within each expedition.

**local\_sample\_id** This isn’t very formally defined. Each expedition had some sort of system for generating unique ids for tissue compartments separated from a given colony. I tried to preserve those here, so that we could find those bags in the freezer.

**tissue\_compartment** One letter code for tissue compartment. In practice, we have this in there as a letter to make it easy to generate the sample ids using CONCATENATE() in Excel.

Letter codes are: M = 'mucus', S='skeleton',T='tissue', W = 'whole' (for outgroups). All others, including Water, etc, ‘None’.

**sample\_type** This is an EMP field, so we tried to consider entries from the point of view of analyzing with other, diverse, EMP samples. Most entries are: Coral Tissue, Coral Mucus, Coral Skeleton. Whole Coral is used occasionally. Water entries are generally ‘Reef Water’ or ‘Water’ (if offshore in ‘blue water’ like some deep buoys we visited for outgroup samples), and most sediment samples are ‘Sediment’. There are some inconsistencies here like ‘Reef Sediment’...we’ve been working on these in more recent versions. Other outgroup samples were more challenging to name, but I usually tried to go with a recognizable, reasonably general outgroup taxon name and then the tissue compartment.

**BiologicalMatter** We treated this identically to sample\_type, but put ‘None’ instead of a categorical value for non-biological samples.

**date** Date in American mm/dd/yy format. **Note:** careful about Excel autoformatting on this one. It does all sorts of weird stuff with text it recognizes as dates, because it internally represents them as integers. Concatenation is particularly obnoxious.

**concatenated\_date** Date as a single integer, always of eight digits of the form yearmonthday. Includes *a zero before single digit days/months. e.g., 08.* *Example: 20150817*

**collection\_time** Collection time in hh:mm format.

**collection\_timestamp.** This field is required for EBI submission. Format taken from the QIITA documentation. mm/dd/yy hh:mm.

**field\_host\_genus\_id**. Genus of the coral, as identified in the field or otherwise listed on sample tubes, notes, etc. None for water or sediment.

**field\_host\_species\_id.** Species of the coral, as identified in the field or otherwise listed on sample tubes, notes, etc. None for water or sediment. *For this column only*, we used sp instead of ‘Unknown’ for unidentified species, so that the result looks like *Porites sp.* when concatenated with the genus, rather than Porites Unknown.

**daily\_replicate.** For GCMP samples, we collect replicates of each coral species in a day. We thought it could be useful to track this.

**binary\_biological.** ‘y’ if the sample is biological in nature, ‘n’ if it is not (e.g. water or sediment).

**field\_host\_clade\_sensu\_fukami\_numeric**  We mapped the field identified genus name to its clade number in the phylogeny of [Fukami 2008](http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0003222). The column has ‘numeric’ in it because clades in the papers are usually given roman numeral names. This looks fancy, but in some cases this is incovenient, so we somewhat redudantly recorded both. When we couldn’t map to clades directly from the paper, we used other literature or consulted with Danwei Huang, a coral systematist who is collaborating on the project.

When comparing clades we thought we’d usually want to compare outgroups and non-biological samples as well, so we put in ‘Water’ and ‘Sediment’. Outgroups we designated with Outgroup, and then their taxonomy starting at the Phylum level and ending with family. Example: Outgroup\_Cnidaria\_Anthozoa\_Corallimorpharia\_Ricordeidae.

**field\_host\_clade\_sensu\_fukami** We mapped the field identified genus name to its clade number in the phylogeny of [Fukami 2008](http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0003222). The clade is given as an uppercase roman numeral. See above for additional details.

**taxonomy\_string\_with\_clade** This was going to be a column in which we listed the taxonomy string starting at phylum for everything, including corals, to allow better comparisons with outgroup. Family would be replaced with clade.

**country** The name of the country. In cases where multiple spellings or names are possible we want to match the GAZ: country codes. I see EMP metadata with GAZ: countryname indicating the locations are terms in the GAZ ontology. However, that ontology also has more specific terms (e.g. Bocas del Toro rather than Panama. Is it necessary to include the GAZ prefix as long as the term is in the ontology? Right now I am checking that country names are valid in the ontology, and not including more specific region names here. Also leaving off the GAZ: prefix for now.

**site\_name** This column, along with reef\_name involved some moderately subjective choices. Often we’ll visit multiple distinct ecosystems or areas in the same country, and dive on multiple reefs in each. So the site name column is the name of the general area within the country (e.g. ‘Farasan Banks’, a set of reefs in Saudi Arabia’s Red Sea coast), while reef\_name captures the specific reef or other location within that area (e.g. ‘Mau Mau Reef’ a specific reef in the Farasan Banks).

**reef\_name** This column captures the specific reef or other location within the general area defined by site\_name and country (e.g. ‘Mau Mau Reef’ a specific reef in the Farasan Banks). Its worth assuming that for most analyses we’ll concatenate country,site\_name,reef to get a unique set of location ids, but we should still avoid reef names that exactly duplicate other reef names from unrelated locations.

**latitude** The latitude of the sample in *decimal degrees.* This is required by the MIxS format. This proved a lot trickier than you might expect due to varying output from GPS systems, which can also output degree-minute or degree-minute-second in a decimal format (that gives crazy coordinates if treated as decimal). A lightweight web converter for other coordinate systems is available [here](http://www.hamstermap.com/). Its worth checking the validity of coordinates by plotting.

**longitude** The longitude of the sample. As latitude. Required by MIxS.

**TAXON\_ID** The NCBI taxon id of the sample. Required by MIxS and the EMP. In some instances no id was assigned for certain coral species. In those cases we used the genus taxon id.

**NCBI\_inherited\_blast\_name** If you look up the NCBI taxon id on the NCBI taxonomy page you also get an ‘inherited BLAST name’. This is the common name for the taxonomic group, which may be useful for broad comparisons across the EMP, where no one would recognize all the species and genus names for such diverse groups.

**TITLE** The title of the project/paper for which the sample was collected.

**reef\_type** The reef type. This is a bit subjective and location specific, but generally captures the role of the reef in the reef system (crest, backreef), and a qualitative description of the distance from any nearby landmass (inshore, midshelf, offshore) etc.

**env\_biome** This is required by MiXS/EMP. It also needs to match a specific term in the EnvO ontology. For 95% of samples ‘coral reef’ was the best term I could fine. See [here](http://bioportal.bioontology.org/ontologies/ENVO?p=classes&conceptid=root)**.**

I am unsure whether the specific ENVO id is needed (e.g. a ‘coral reef’ is ENVO 00000150). I’d prefer to exclude and just use ENVO compatible terms if at all possible since these are much less verbose (and it should be possible to uniquely look up the id from ENVO by the term (since that’s part of the point of having a controlled vocabulary).

**env\_feature** This is required by MiXS/EMP. It also needs to match a specific term in the EnvO ontology. See the environmental feature section, browsable [here](http://bioportal.bioontology.org/ontologies/ENVO?p=classes&conceptid=root)**.**

**env\_matter** EnvO doesn’t currently have all the entries we’d want for types of matter. For example, they have mucus, but not skeleton or tissue (much less aragonite skeleton). So mucus is ‘mucus’ and skeleton and tissue are ‘organic material’. Ideally, we should get EnvO to add the following terms:

Cnidarian-associated habitat. is\_a animal\_associated\_habitat

Coral-associated habitat is\_a cnidarian-associated habitat

has\_a surface mucus layer (environmental matter)

has\_a tissue

has\_a symbiosome

has a aragonite skeleton

has\_a gastric cavity

tissue

child\_of organic material

**substrate** The substrate the coral is sitting on. Trying to record differences between corals sitting in sand vs. on reef walls for example. This column is recorded somewhat inconsistently in this version (we’ve worked out some ideas for revising).

**depth** Depth at which the sample was collected, in meters.

**temperature** Temperature at depth where the coral was collected in celsius.

**temperature\_method**  Records the method by which temperature was collected. Common values are: ‘dive\_computer’, ‘HOBO\_logger’,’Seabird\_logger’,’Remote\_sensing’

**surface\_temperature** Temperature at the surface where the coral was collected, if known.

**salinity** If measured. Only recorded for a few samples. Need to standardize units.

**oxygen** If measured. Only recorded for a few samples. Need to standardize units.

**photosynthetically\_active\_radiation** If measured. Only recorded for a few samples. Need to standardize.

**PAR\_method** Method for measuring PAR (e.g. instrument at the coral

**visibility**

**dna\_extracted** ‘y’ if DNA is extracted. Note that EMP requires MoBio PowerSoil extraction. Many GCMP samples will be updated as a final step, since extractions are in progress.

**physical\_samp\_avail\_now** ‘y’ if a physical sample (i.e. frozen) is available. ‘n’ if its just extracted dna.

**physical\_sample\_location.** Location of physical specimen, if available, and/or DNA. Free text. Required by EBI. Bizarrely, the EMP asks for no longer available samples to have entries listed as ‘the location, if a sample *were* available’. This is too existential for me, so I suggest that any samples without physical specimens **or** DNA available be excluded. (If DNA but not the physical specimen is available we list the location of the DNA).

**colony\_height\_cm** Height of the colony (i.e. vertical size) in centimeters.

**colony\_width1** Largest width dimension of the colony (centimeters)

**colony\_width2** Smaller width dimension of the colony (centimeters)

**coral\_color** This is the coral color from the CoralWatch color cards.

**coral\_color\_additional** Sometimes corals had two distinct colors, or an unusual color that might reflect disease, etc in small portions of the colony. These values were recorded here when present as free text.

**binary\_tissue\_loss** ‘y’ if the coral appears to have lost tissue, ‘n’ otherwise

**tissue\_loss\_percent** if the coral appears to have lost tissue, divers estimated the percent. This method is inexact, but we think it will be more robust than categories, since for example 0-25 biases a lot of low percent estimates upward when converted to a percentage for regression. Ideally all of these would be measured in the field, but in practice exact measurements weren’t feasible given time constraints in many areas. We may try quantifying a subset of these in ImageJ from the Flickr photos to be more quantitative.

**disease\_name** The name of any coral disease present. If unknown, a descriptive name can be used.

**binary\_disease** ‘y’ if the coral appears to have a disease, ‘n’ otherwise.

**disease\_percent** estimated percent of the coral affected by the disease.

**cyanobacteria\_contact** ‘y’ if the coral is contacting visible cyanobacteria, ‘n’ otherwise.

**cyanobacteria\_percent** estimated percentage of the perimeter of the coral contacting cyanobacteria. If internal patches of dead tissue also have cyanobacteria, count them as part of the overall perimeter.

**binary\_turf\_contact** ‘y’ if the coral is contacting turf algae, ‘n’ otherwise.

**turf\_contact\_percent** estimated percentage of the perimeter of the coral contacting turf algae. If internal patches of dead tissue also have turf, or the coral is partially overgrown, include overgrown or internal patches as part of the overall perimeter.

**binary\_algal\_contact** ‘y’ if the coral is in contact with either turf algae or macroalgae, ‘n’ otherwise. Note that as CCA isn’t considered harmful (and may be beneficial) we don’t include it here.

**n\_algal\_contacts** Number of distinct types of macroalgae and/or turf algae contacting a coral.

**algal\_contact\_types** A list of the algal contact types, including turf and CCA. **Note:** since some algae have multi-word names/descriptions, I’ve been separating algal types with two underscores ‘\_\_’ so we can split these back out easily later. Try to see if one of the existing values matches before entering a new one (its better to not have ‘Halimeda\_\_Turf’ and ‘Turf\_\_Halimeda’ for example.

**binary\_macroalgal\_contact** ‘y’ if contacting macroalgae, ‘n’ otherwise.

**n\_macroalgal\_contacts** Number of distinct types of macroalgae contacted (even if they can’t all be fully identified). Doesn’t include turf algae/CCA.

**macroalgal\_contact\_types** A list of the algal contact types, excluding turf and CCA. **Note:** since some algae have multi-word names/descriptions, I’ve been separating algal types with two underscores ‘\_\_’ so we can split these back out easily later. Try to see if one of the existing values matches before entering a new one (its better to not have ‘Halimeda\_\_Dictyota’ and ‘Dictyota\_\_Halimeda’ for example.

**binary\_CCA\_contact** ‘y’ if the coral is contacting crustose coralline algae, ‘n’ otherwise.

**cca\_contact\_percent** estimated perimeter of the coral touching crustose coralline algae.

**seagrass\_contact** ‘y’ if the coral is touching seagrass, ‘n’ otherwise

**parrotfish\_bites** ‘y’ if the coral has identifiable parrotfish bite scars, ‘n’ otherwise. **Note:** we may need to expand this to include other types of corallivory that leave distinctive markings. We can revisit as samples that include these features emerge.

**contact\_description** A straight copy of the description of contacts from the dive notes. Useful in case we make new categories for things like contact with wreckage, trash and metal bars for example

**invertebrate\_contact\_types** Types of benthic invertebrates contacted by the coral, excluding other corals (same format as algal\_contact\_types). Note: not available for many corals.

**coral\_contact\_types** Types of other corals contacted by the coral. Note: not available for many corals.

**Sediment\_contact** ‘y’ if sediment is deposited on top of the coral or the coral was partially buried.

**Description** Required by QIIME. We used an automatically generated description of the sample. *Example:*

|  |
| --- |
| Coral Mucus from Echinopora mammiformis sampled from Great Barrier Reef - Northern Sector, Australia as sample E3.17.Ech.mamm.L45.20150129.M of the Global Coral Microbiome Project |

Note that QIIME requires descriptions for every sample, and they have to be unique.

Columns pertaining only to the Kiritimati Reef Resilience Project samples:

**sediment\_volume**: Volume of sediment sample (mls)

**water\_volume:** Volume of water sample (mls)

**Extracted\_Chl:** Extracted chlorophyll value at the surface water (average of three per site)

**chlorophyll\_insitu:** Surface water chlorophyll measurements from Aquafluor instrument, (average of three samples per site)

**turbidity:** Surface water turbidity measurements from Aquafluor instrument (average of three samples per site)

**Rain:** If there was rain during the sampling of surface water

**Wind:** If there was wind during the sampling of surface water

**nitrate\_nitrile**: Surface water measurements of nitrite (uM) (average of three per site)

**phosphate**: Surface water measurements of phosphate (uM) (average of three per site)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |

**Temp\_insitu\_SBE56\_Max**: Maximum water temperature for the month before the sampling period (same depth as our coral transects) sampled by a Seabird 56 temperature logger

**Temp\_insitu\_SBE56\_Min:** Minimum water temperature for the month before the sampling period (same depth as our coral transects) sampled by a Seabird 56 temperature logger

**Temp\_insitu\_SBE56\_Min:** Mean water temperature for the month before the sampling period (same depth as our coral transects) sampled by a Seabird 56 temperature logger

**Temp\_insitu\_HOBO\_Max:** Maximum water temperature for the month before the sampling period (same depth as our coral transects) sampled by a HOBO temperature logger

**Temp\_insitu\_HOBO\_Min:** Minimum water temperature for the month before the sampling period (same depth as our coral transects) sampled by a HOBO temperature logger

**Temp\_insitu\_HOBO\_Mean:** Mean water temperature for the month before the sampling period (same depth as our coral transects) sampled by a HOBO temperature logger

**Remote Sensing Variables**

**RS\_sstmax:** average maximum sea surface temperature

**RS\_sstmean\_RS:** mean sea surface temperature

**RS\_sstmin:** average minimum sea surface temperature

**RS\_East/West Aspect:** East/West aspect

**RS\_North/South Aspect:** North/South aspect

**RS\_Plan Curvature:** Plan curvature

**RS\_Profile Curvature:** Profile curvature

**RS\_Distance to shore:** Distance to shore

**RS\_Bathymetric Slope:** Bathymetric Slope

**RS\_Concavity:** Concavity

**RS\_bathymetry:** bathymetry

**RS\_cortad.derived.avg.ann.severe:** Average annual number of severe bleaching events (average

number of weeks per year with DHW values equal to or greater than 8)

**RS\_cortad.derived.avg.ann.stress:** Average annual bleaching stress

**RS\_cortad.derived.avg.max.sst:** Average maximum SST

**RS\_cortad.derived.avg.min.sst:** Average minimum SST

**RS\_cortad.derived.avg.range.sst:** Average intra-annual temperature range

**RS\_cortad.derived.avg.sd.sst:** Average intra-annual temperature variability

**RS\_cortad.derived.avg.severe.duration:** Average duration of severe bleaching events (average

number of consecutive DHW values of 8 or greater)

**RS\_cortad.derived.avg.severe.interval:** Average interval between severe bleaching events (average interval between DHW values of 8 or more)

**RS\_cortad.derived.avg.sst:** Average SST

**RS\_cortad.derived.avg.stress.duration:** Average bleaching stress duration (average duration of bleaching-prone periods)

**RS\_cortad.derived.avg.stress.interval:** Average bleaching stress interval (average interval between bleaching-prone periods)

**RS\_cortad.derived.climate.trend:** Climate trend

**RS\_cortad.derived.cumfreq.coldwater:** Cumulative frequency of cold water events

**RS\_cortad.derived.cumfreq.hotwater:** Cumulative frequency of hot water events

**RS\_cortad.derived.cumfreq.stheat:** Cumulative frequency of short-term heat stress for fish

**RS\_cortad.derived.sd.mean.sst:** Inter-annual temperature variability

**RS\_cortad.derived.tot.severe:** Cumulative number of severe bleaching events (number of weeks with DHW values equal or greater than 8)

**RS\_cortad.derived.tot.stress:** Cumulative bleaching stress

**RS\_reef.area.millennium.km.sq:** Reef area (the total reef area that is contiguous with the survey site)

**RS\_closest.reef.dist:** Reef isolation- distance to the next nearest reef (i.e. island/atoll)

**RS\_area\_75km\_buffer\_sq\_km:** Reef isolation- area based (total non-reef area within 75km)

**RS\_area\_300km\_buffer\_sq\_km:** Reef isolation- area based (total non-reef area within 300km)

**RS\_calcite:** Calcite

**RS\_chlomax:** Chlorophyll max

**RS\_chlomean:** Chlorophyll mean

**RS\_chlomin:** Chlorophyll mean

**RS\_chlorange:** Chlorophyll range

**RS\_cloudmax:** Cloud max

**RS\_cloudmean:** Cloud mean

**RS\_cloudmin:** Cloud min

**RS\_damax:** Diffused attentuation max

**RS\_damean:** Diffused attentuation mean

**RS\_damin:** Diffused attentuation min

**RS\_dissox:** Dissolved oxygen

**RS\_nitrate:** Nitrate

**RS\_parmax:** Photosynthetically active radiation max

**RS\_parmean:** Photosynthetically active radiation mean

**RS\_pH:** pH

**RS\_phos:** Phosphorous

**RS\_salinity:** Salinity

**RS\_silicate:** Silicate

**RS\_sstmax:** SST max

**RS\_sstmean:** SST mean

**RS\_sstmin:** SST min

**RS\_sstrange:** SST range

**RS\_arag\_380:** Global threat-current arogonite saturation

**RS\_therm\_98-07:** Global threat- past thermal stress

**RS\_thr\_cd:** Local threat- coastal development

**RS\_thr\_int\_loc:** Integrated local threat- combining all four local threats

**RS\_thr\_mar:** Local threat- marine based pollution

**RS\_thr\_ovf\_adj:** Local threat- overfishing/destructive fishing adjusted for MPA management

**RS\_thr\_sed:** Local threat- watershed based pollution