NCRMP Sample Frame Development Protocols for U.S. Caribbean and FGBNMS

National Coral Reef Monitoring Program
Coral Reef Conservation Program (CRCP), National Oceanic and Atmospheric Administration

The following workflow outlines the protocols developed from 2013-2018 to design and generate a gridded, stratified sample frame for particular study regions monitored by NOAA's NOS/NCCOS/Biogeography Branch and its partners as part of the National Coral Reef Monitoring Program (NCRMP) directed by NOAA's Coral Reef Conservation Program (CRCP). A stratified random sample selection (using weighted strata allocations) is applied to a region's sample frame to select sites for each survey mission. These regions are surveyed biennially, and each sample frame is updated or re-created prior to each survey mission to incorporate new strata data or reflect protocol changes as they have evolved and improved over the years. The following workflow is up to date as of 2018 and specifically addresses the methodology of sample frame generation. Broader protocols for sample design and allocation can be found in a separate document.

Note about Flower Gardens: much of these steps apply only to USVI and Puerto Rico sample frames. The FGB frame is much smaller and generally static from year to year without strata updates. FGB habitat type is much simpler and the hab allocation process described below is not applied to FGB. Bathymetry is relatively high resolution with good coverage over diveable areas, and has not been updated in recent years. As a result, site selection from the sample frame is often the only step required for FGB.

Note about Puerto Rico: There is 1 base grid for STJ&STT, 1 base grid for STX, and 1 base grid for FGB. Due to its large size and processing demands in ArcGIS, the base grid for PR is split into four quadrants (cells are assigned to quadrants by centroid location). These are developed separately into four sample frames, i.e. all steps below repeated 4x, and the quadrants are merged as the final step. It is likely this splitting can be avoided in a scripting environment in the future.

File Storage:

Sample frame data are currently stored on the NOAA/NOS/NCCOS/CCMA network "T:" drive. All rasters and shapefiles used in sample frame creation are stored in a file GDB used to house GIS data for a given mission, named [REGION]_[YEAR]_Sample_Design_Package.gdb. This GDB includes files such as bathy and habitat maps used to develop a sample frame, several iterative products generated at different points in the development process, and the final sample frame and site selection subset used for the mission.

All sample design components/products must have the appropriate projection: UTM19N for PR, UTM20N for USVI, and UTM15N for FGB.

Two feature datasets are created in the GDB: DesignCriteria, which should hold final strata layers such as admin areas and habitat, and SampleDesign, which holds centroids and grids.

The workspace for housing intermediary files and drafts for a given mission is here:
T:\Projects\NCRMP\[YEAR] Missions\[REGION] [YEAR]\Sample Design\Sample Design Development
The GDB of finalized files and accompanying MXD are here:
T:\Projects\NCRMP\[YEAR] Missions\[REGION] [YEAR]\Sample Design\Final Sample Design
E.g., "T:\Projects\NCRMP\2015 Missions\FGBNMS 2015\Sample Design\Final Sample Design"

Geographic scope and base grid determination:

Scope is defined as all areas on the insular shelves and within US waters. For St John and St Thomas the geographic scope extends to longitude -65.15 on the western boundary, where the areal extent abuts the Puerto Rico sample frame. Base grids are regularly spaced 50 m x 50 m cells which serve as the primary sampling units. All cells are assigned unique IDs which are preserved throughout sample frame development iterations and serve as identifiers for individual cells for all GIS analysis (final subset of selected sites are receive new cell names for field planning).

Biotopes Development

Biotopes (broad ecological zones) were defined and drawn by Biogeo staff into shapefiles dividing the study area into major zones for sampling. Their outer borders define the spatial extent of the sample area. These are defined for USVI and PR and should not require annual updates. Every cell in the sample frame falls within one biotope and that Biotope stratum is assigned to it. The Biotope shapefile is combined with the Admin Areas shapefile (below).

In USVI, units were also classified into one of the eight following biotopes: 1) MSR – the mid-shelf reefs off St. Thomas and St. John, 2) SLRK – the Sail Rock area, 3) STJ – nearshore St. John, 4) STT – nearshore St. Thomas, 5) LANG – the Lang Bank area, 6) ESTX – the eastern side of St Croix consisting of BUCK and EEMP, 7) NWSTC – the northern and western sides of St Croix, 8) SSTC – the southern side of St Croix. The precise boundaries of biotopes attempt to overlap distinctions in administrative zones, depth classes and benthic habitat types.
In Puerto Rico, units were classified into one of the three following biotopes: 1) N – northern area of Puerto Rico, 2) E – eastern area of Puerto Rico, and 3) SW – southwestern area of Puerto Rico.

Administrative Areas Development

Hopefully, all admin areas (e.g. protected areas, marine parks, etc.) have been clearly defined and the appropriate shapefiles are complete with no need for annual update. More closures may be added to the compilation if they become important enough to our partners to become a new stratum. Any new shapefiles to be added, such as a new MPA or revised version of a previously included one, likely come from external sources. They should be visually inspected in Arc for border gaps and other drawing errors. Some closed areas required manual editing to tidy them up as a last resort. If the area is in multiple segments, they should be dissolved together into one feature.

In the USVI, units were classified into one of the seven following Administrative areas independent of the Biotope selection process: (1) VICR – the Virgin Islands Coral Reef National Monument, (2) VINP – the Virgin Islands National Park, (3) STEER – the St. Thomas East End Reserve, (4) BUCK – The Buck Island Reef National Monument, (5) EEMP – the East End Marine Park, (6) SARI – Salt River Bay National Historic Park and Ecological Preserve, and (7) OPEN – the remaining units.

In Puerto Rico, units were classified into one of the five following Administrative areas: (1) CLP – the Canal Luis Peña NR, (2) ID – the Isla de Desecheo Marine Reserve, (3) IM – the Isla de Mona Natural Reserve, (4) TPR - Tres Palmas de Rincón Marine Reserve, and (5) OPEN – the remaining units. There are no Admin areas in FGBNMS.

Admin areas and Biotopes should be merged together in Arc to form one AREAS shapefile. Barring any changes, the same AREAs file from previous years will be used. Admin areas will overlap and be contained within one or more of the larger Biotope areas. In the AREAS shapefile, BIOTOPE and ADMIN are two separate fields and both should be populated for all features. For all "unprotected" Biotope areas where there are no designated admin areas, populate as ADMIN = OPEN. Every grid cell will ultimately be assigned to one Biotope and one Admin Area stratum value.

The merged Areas should be added to the GDB/DesignCriteria dataset.

Base Grid Development:

The Base Grid is created with the ArcGIS Fishnet tool, projected appropriately and constrained in size. A well-defined base grid may exist and not need to be recreated unless the following size constraint factors have been updated since the last mission.

The first step to constrain an oversize base grid is to reduce it to the borders of the Biotope zones (AREAS feature) created for the study region. Use Select by Location tool to extract only those cells with centroids inside the borders of AREAS.

Second, shallow-water subset of the region's Coastal Relief Model should be made where depth >= -100 m. Use Select by Location to extract only those cells within the subset bathy.

Third, remove grid cells from known land areas. A land shapefile is generated by exporting two hab types, "Land" & "Mangrove", from the final habitat map used for the sample frame (see development steps below) and exporting to a new shapefile LAND. A 5 m buffer is added to the land feature, and all grid cells with their centroids outside the LAND_5m_BUFFER feature are selected and exported to a new base grid with land areas removed (using Select by Location tool). This land buffer feature already exists for Caribbean study areas and may not need to be recreated.

These initial reductions reduce processing time. Base grids have already been created in this fashion for all Caribbean regions and this initial base grid should not need updates unless Admin Area polygon or the CRM is revised, which is unlikely. The Base Grid is further reduced in later iterations by additional depth sources and habitat, but these may change each year with new data.

Base grid constrained by Admin Area boundaries and a 100 m depth max using the CRM should be included in the GDB/SampleDesign and have following file attributes:

BASE_GRID – grid cells (fishnet) without stratification information.
Filetype = shapefile, location =Sample Design Development folder
Attributes:
GRID_ID - Unique identifier
REGION - group set {PRICO, STX, STTSTJ, FGB}

In the case of Puerto Rico, after this base grid is created, it is split into 4 quadrants (BASE_GRID_NE, BASE_GRID_NW, etc.) added to the GDB and proceed with steps below for each quadrant.

Bathymetry Data Compilation

A running compilation of all useful bathymetry overlapping the region's study area is maintained and updated each year with new datasets as they become available. Only those bathy datasets used in the sample frame are kept in the GDB. All datasets must be projected appropriately and depth values formatted to be negative (-). The "base" or "last resort" bathy source for most regions is the CRM (1arc second or ~30 m cell size) from NOAA's NCEI, which has the biggest geographic scope but generally the poorest resolution and most uncertainty. A similar background bathy source are the Coastal Digital Elevation Models, also interpolated surfaces from NCEI, but smaller in size and higher in resolution (10 m). Whenever possible however, higher-resolution multibeam or Lidar datasets are used, typically obtained from NOAA NCCOS, NCEI, NGS or USGS data portals.

All project bathy layers to be included in the sample frame should be individually exported to the GDB. (Rasters cannot be put into Feature Datasets so there is no subdirectory for them).

Bathymetry Incorporation

Bathy datasets are not merged into a single bathy surface due to widely varying resolutions. Instead, they are kept separate and ranked hierarchically based on resolution (and sometimes quality, based on communication with data authors).

Centroids are exported from the base grid as a pointfile where GRID_ID is preserved for each point. Depth is assigned to each point using the best available bathy source occurring at that point according to the ranking hierarchy. In the past this hierarchical selection was accomplished by using Extract Values to Points tool for each bathy raster used in the region, adding a new field for each raster in the centroid file (given the small size of most bathy datasets compared to the base grid, there are 0 values for many points and bathy fields). The best available bathy source is selected hierarchically for each point using if/else/then statements in ArcPy and that 'best' depth value is assigned to a new field, DEPTH_M. Depth values in this field should be negative and field type should be "Double."

Two depth classes are used to stratify samples: less than 12 meters, and greater than or equal to 12 meters and less than or equal to 30 meters. Another categorical field, DEPTH, is populated for each centroid to indicate depth category based on the numerical DEPTH_M value using if/then/else statement in ArcPy.
If DEPTH_M>=0, then DEPTH = LAND
If 0>DEPTH_M> -12, then DEPTH = SHLW
If -12 >=DEPTH_M>= -30, then DEPTH= DEEP
If DEPTH_M< -30, then DEPTH = GTR30
Points outside that depth range should be classified as Land or GTR30 but retained at this stage.

Latitude and longitude (decimal degree format) are calculated for each centroid in new fields using Calculate Field Geometry in the attribute table. A copy of this point file should be made

and the various bathy fields from different rasters deleted except for the final DEPTH_M and DEPTH fields.

The centroid file should be included in the GDB/SampleDesign dataset with the following attributes:

INTERSECT_CENTROIDS - Centroids from intersect grid. Needed for processing.
Filetype = shapefile, location =Sample Design Development folder
Attributes:
GRID_ID – Unique identifier
REGION - group set {PRICO, STX, STTSTJ, FGB}
LAT – latitude in decimal degrees (field type Double)
LONG – longitude in decimal degrees (field type Double)
DEPTH_M – depth in meters (should be negative if below surface, field type Double)
DEPTH - group set {SHLW, DEEP, GTR30, LAND}

Finally, the attributes of these centroids should be added to the base grid using Spatial Join, matching each centroid to the grill cell it is inside. Join Field (matching by GRID_ID) also works, but seems to be slower in Arc. Spatial join will produce a new grid, to be named as a second, intermediary version of the Base Grid. It will be populated with additional attributes and another version generated.

Habitat Map Creation

All benthic habitat data available from US government agencies is merged using a hierarchy dependent on coverage, map accuracy and spatial resolution. Habitat data are updated before every mission. Five benthic habitat types are used to stratify samples: scattered coral and rock in sand, bedrock, pavement, patch reef, aggregate reef. A set of rules has been established to reclassify other habitat types in some maps into these five categories. Where benthic habitat class maps are missing in the study area (typically in deeper water), predicted hardbottom (aka hardbottom proxy) habitat maps are used instead, which lack habitat classification other than a differentiation of soft and hard bottom based on bathymetry interpolation. These areas of Predicted Hardbottom are added as a sixth hard habitat class. Predicted hardbottom has been updated several times since NCRMP began, and Matt Poti at NOAA/NCCOS Biogeography Branch should be consulted about future updates.

Some habitat maps have polygons classified as Unknown, and these must be deleted from the shapefile before further analysis. Next, where different habitat maps spatially overlap, the map with lower resolution/accuracy is Erased from that area of overlap. Conducting these Erases at all point of overlap produce a patchwork of contiguous but non-overlapping habitat maps that are then Merged into a unified map. Similarly, where Predicted Hardbottom polygons overlap known hard habitat classes, Predicted Hardbottom is erased, and Predicted Hardbottom polygons that don't overlap known hab maps are merged with the final habitat map. Given different field names and attributes of different maps, the attribute table of the final merged

habitat map must be heavily edited and values moved between fields. Hardbottom habitats for sampling consist of the following types: aggregate or linear reefs, patch reefs, colonized bedrock, colonized pavement, scattered coral/rock in unconsolidated sediments and predicted hardbottom. Reef rubble, rhodolith habitats, and artificial reefs are not included as hardbottom habitats.

A numerical field, HAB_CODE, is created to assign an integer value for each hab type for later processing and condense non-hardbottom habitat types. The field should be populated as follows:
HAB_CODE (Integer):
Scattered coral & rock = 5
Pavement = 6
Bedrock = 7
Patch reef = 8
Aggregate reef = 9
Predicted Hardbottom = 10
All soft bottom or hab types that will not be sampled = 999

The final merged habitat map is added to the GDB/DesignCriteria.

Habitat Incorporation

The Base Grid should be reduced in size to where it overlaps the habitat map (Select by Location all cells that intersect hab map and Export as new file). The habitat map and reduced grid are combined by Union tool, producing a grid of cells split along habitat lines. Hab data is assigned to each cell and where multiple habitat are present in the cell, multiple polygons are produced with different habitat but the same cell ID.
The Hab Union should be Dissolved by GRID_ID and HAB_CODE, producing Hab_Union_Dissolve shapefile. This ensures that if multiple polygons are present in the same cell that have the same Hab Code, they will be considered one feature and their areas will be combined.

Two fields are manually added to resulting shapefile:
HAB_AREA: Calculate Field Geometry in attribute table to populate with area in square meters. Provides area of hab type within each cell. Hab type at this stage is based on the numerical code, and all non-hardbottom (999) habitats have been dissolved into one feature and one area per cell.
PERC_AREA: =(HAB_AREA/2500)*100. Gives % of total cell area that a given hab type comprises. 2500 m^2 is the total area of each cell.

The majority of grid cells have the same habitat code for the entirety of the cell. For those with multiple hab types, however, one hab type must be selected and assigned to the cell, and this is accomplished using a habitat allocation algorithm written as a Python script. The

Hab_Union_Dissolve shapefile is exported as a DBF and run through the script, which evaluates each cell by a few simple rules:

1. If percent area of 999 habitat is >90% for a given, that cell is assigned as 999 (not hard habitat and will not be surveyed).
2. If <90%, this means at least 10% of the cell is some type of hard habitat. A minimum of 10% hardbottom habitat is deemed sufficient to be surveyed and included in the NCRMP sample frame. If only one hardbottom type is present, that is the hab type assigned to the cell.
3. If multiple hard hab types are present and when combined, amount >10% of the cell area, whichever hard hab type has the greatest area will be assigned to the cell. (This means that although extremely rare, it is theoretically possible that the hard hab type assigned to the cell comprises <10% area, if there are other hard hab types present that push the collective sum over the 10% mark).

The output is written as a new text file where there is only one Hab Code for each cell ID (i.e. entry per cell ID). The PERC_AREA field is also preserved to allow users to see what the actual area coverage for that hab type is, in case there are questions about the site during or after the mission. Convert text file to DBF.

Using Join by Field in Arc (matching based on GRID_ID), join new DBF file attributes to Intersect Grid, including HAB_CODE & PERC_AREA.

In the Intersect Grid, revised HAB_CODE into a text label for habitat type. Add new text field HABITAT and populate as follows:
If HAB_CODE = 5 → HABITAT = SCR
If HAB_CODE = 6 → HABITAT = PVMT
If HAB_CODE = 7 → HABITAT = BDRK
If HAB_CODE = 8 → HABITAT = PTRF
If HAB_CODE = 9 → HABITAT = AGRF
If HAB_CODE = 10 → HABITAT = HARD
If HAB_CODE = 999 → HABITAT = OTHR

For FGB, habitat has already been determined in previous years for each cell, and the only two HABITAT values are "HRLF" and "LRLF" (high and low relief reef).

Biotope and Admin Incorporation

Conduct spatial join on AREAS and to create new intermediate grid. called INTERSECT_GRID. with BIOTOPE and ADMIN fields added to all cells. Two integer fields, HZRD_FLAG and SHOR_FLAG, should be added and left blank for now. They will be used to mark problematic cells in subsequent steps. All major attributes should now be included in this base grid draft. Although it has been constrained to cells within AREAS border, the grid still includes many cells with depths outside the 0-30 m sample range, and cells classified as OTHR (not hard bottom).The grid attributes should be as follows:

INTERSECT_GRID – base grid stratified by many layers.
Filetype = shapefile, location = Sample Design Development folder
Attributes:
GRID_ID – Unique identifier
REGION - group set {PRICO, STX, STTSTJ, FGB}
LAT – latitude (taken from centroid)
LONG – longitude (taken from centroid)
DEPTH_M – depth in meters (taken from centroid)
PERC_AREA – percent area of majority habitat class from intersect table
DEPTH – group set (taken from centroid) {SHLW, DEEP, GTR30, LAND}
HABITAT - class from intersect table {HARD, BDRK, SCR, AGRF, PTRF, PVMT, OTHR, HRLF, LRLF}
BIOTOPE – group set (taken from centroid) {STJ, STT, MSR, SLRK, EDGE, ESTX, NWSTX, SSTX, LANG, EPRICO, NPRICO, SWPRICO, EFGB, WFGB}
ADMIN – group set (taken from centroid) {OPEN ,VIIS, VICR, STEER, SARI, BUIS, EEMP, MONA, DESE, TPR, CLP}
HZRD_FLAG – boolean set (taken from centroid) {0,1}
SHOR_FLAG – boolean set (taken from centroid) {0,1}

Add grid to GDB/SampleDesign.

Edit Intersect Grid, Remove Cells from Frame

A copy of Intersect Grid is made next and named SAMPLE_FRAME_DRAFT, and this copy will reduced through many steps. Conduct a series of checks and subsequent deletions, often using the Select by Location tool.

1. Check that there are no cells with centroids outside of the outer borders of the AREAS polygons.
2. If HABITAT = OTHR, cells are deleted.
3. If DEPTH = LAND or GTR30, cells are deleted.
4. All cells with centroids inside LAND_5m_BUFF feature are deleted.
5. There should be no DEPTH_M values that are positive (on land), or deeper than -30.
6. There should be no blank values in any field. Each field should be checked for missing values, and these cells should be investigated. Values should be filled in if possible or the cells should be deleted, depending on specifics.

Shore & Hazard Flagged Site Removal

The HZRD_FLAG and SHOR_FLAG fields should be populated where necessary for this Sample Frame draft after above cells are removed. This requires visually inspecting the

shoreline of all land masses in the study region using the best available aerial photography in ArcGIS. Inspection should be thorough and methodical. This may have already been done in previous years, and if so that file should be overlaid on the sample frame draft (see below). If not: for each cell that appears to have its center located on land in the photography image, or where depth appears to be zero (often where sand and surf meet, or waves are breaking on a rock), flag that cell by typing manually in an Edit session in Arc. If flagged for being on the shoreline, then SHOR_FLAG=1. If no flag, SHOR_FLAG field remains blank. If there is any doubt about whether the cell center might be too shallow to dive, flag it. Better to remove it from the sample frame than waste time on a problematic site in the field.

There may be some small ADMIN areas that actually define areas of exclusion that should not be sampled, such as a major ship transit passage in St. Thomas. Cells w/ centroids within these areas should be flagged. For these sites, HZRD_FLAG ="1."

The shoreline should be inspected to identify any additional cells that appear to be in hazardous areas for diving, primarily shipping channels and around marinas. Local partners can be consulted to identify those sites and how much of a buffer there should be between them and dive ops. These sites should be flagged in the same way: HZRD_FLAG=1 if the site is considered potentially hazardous.

When the entire shoreline and every small island has been evaluated and all cells flagged, every flagged cell (both shoreline and hazard) should be selected and exported as a subset shapefile. This FLAGGED_SITES shapefile can be used to quickly identify nearshore sites for removal for the next mission. One may already exist and can be used in place of repeating the entire inspection process, if there are not major updates to the frame that require it to be done again from scratch. If already present, overlay the flagged sites subset over the new sample frame draft. Use Select by Location to ID the sample frame sites that overlap the flagged areas and delete them.

Final Sample Frame Products

A final copy of this sample frame draft should be created, and all flagged sites removed. Add a text field called STRATUM to final sample frame. Populate by concatenating all of the strata fields into one stratum (see sequence below).

For all cells where Habitat= BDRK, DEPTH should be changed to SHLW, e.g. change DEEPBDRKSTTOPEN to SHLWBDRKSTTOPEN. This is to condense strata due to the scarcity of deep Bedrock sites and the likelihood that a DEEP BDRK site is only 1-2 meters deeper than the 12 m division.

For Puerto Rico, at this point merge all four quadrants into single final sample frame. Final sample frame for all regions should be structured as follows and added to GDB/SampleDesign:

SAMPLE_FRAME – Subset of intersect grid within all constraints.
Filetype = shapefile, location =Sample Design Development folder
Attributes:
same as INTERSECT_GRID with addition of:
STRATUM - stratum identifier {DEPTH & HABITAT & BIOTOPE & ADMIN}

Use sample allocation scripts/tools on Sample Frame to determine number of sites to allocate to each stratum, and run site selection. Processes should generate the following shapefile and CSV table to include in the GDB/SampleDesign:

SAMPLE_SELECT – Subset of sample frame using allocation procedure.
Filetype = text file and shapefile, location =Sample Design Development folder
Attributes:
same as SAMPLE_FRAME with addition of:
TYPE - identifier for primary and auxiliary sites {primary, auxiliary}

STRATUM_TABLE – Summary of strata selection data
Filetype = text file, location =Sample Design Development folder
Attributes:
STRATUM - stratum identifier {DEPTH & HABITAT & BIOTOPE & ADMIN}
DEPTH - group set {SHLW, DEEP, GTR30}
BIOTOPE – group set {STJ, STT, MSR, SLRK, EDGE, ESTX, NWSTX, SSTX, LANG, EPRICO, NPRICO, SWPRICO}
ADMIN – group set {OPEN ,VIIS, VICR, STEER, SARI, BUIS, EEMP, MONA, DESE, TPR, CLP}
HABITAT - class from intersect table {HARD, BDRK, SCR, AGRF, PTRF, PVMT, OTHR, HRLF, LRLF}
NTOT - number of sample units in stratum
WH - proportion of sample units from sample frame within stratum

HABVAR - habitat variance used in sample allocation equations
NH - number of primary samples selected within each stratum
NH_ADD - number of secondary (extra) samples selected within each stratum
NH_PLOT - number of primary + secondary samples selected within stratum

After the field mission is complete, a sampled site summary text table following this general format should also be added to the Final Sample Design folder:

OBSERVATION_TABLE - summary of post field mission data
region
survey_year
survey_date
survey_index
station_code
latitude
longitude
biotope
habitat
depth
admin
stratum
fish
LPI
demo
N_cells_stratum
N_fish_cells_sampled
fish_sampling_weights
N_LPI_cells_sampled
LPI_sampling_weights
N_demo_cells_sampled
demo_sampling_weights