Project Title: Synthesizing spatial dynamics of recreational fish and fisheries to inform restoration strategies: red drum in the Gulf of Mexico

A Data Synthesis Project supported by the National Academies Gulf Research Program
Total budget requested: $ 480,248

Key contacts
Kai Lorenzen (Project Director, PI), University of Florida (UF), School of Forest Resources and Conservation (SFRC), Fisheries and Aquatic Sciences Program (FAS): klorenzen@ufl.edu, 352-273-3646 (general enquiries)
Ed Camp (PostDoc, Co-PI), UF, SFRC, FAS: edvcamp@ufl.edu (modeling and quantitative analysis)
Jynessa Dutka-Gianelli (PostDoc, Co-PI), UF, SFRC, FAS: jdgianelli@ufl.edu, 352-392-9981 (data management)

Revised Data Management Plan
[Following reviewers' suggestions. Retains header and table numbering from the full proposal document]

d.i Data and other information products: Planning

The project will capture and synthesize multiple types of pre-existing observational data from various sources and formats. The original sources of the anticipated data, as well as pertinent information describing how such data were collected and relevant sharing arrangements are described in Tables 2-4. Information regarding the type and formats of key data sets is given in Table 5. In addition to these key data sets derived from major long-term and large-scale sampling programs, a variety of smaller data sets originating from individual research projects may be captured. Formats of these data sets will be assessed in the course of the project.
We note that observational data (Table 5) are captured in a variety of different formats. Geospatial data, GIS products and remote sensing results from different platforms will be incorporated, following geospatial standards such as Open Geospatial Consortium (OGC) Web Map Service (WMS) to enable spatial data visualization standard software such as ESRI ArcGIS and Google Earth. All image files and geospatial information will be with defined Projection / Coordinate reference system and datum. Digital base maps and diverse datasets available (e.g., physical, biotic, environmental quality layers) will be downloaded from the NOAA National Coastal Data Development Center, Gulf of Mexico Data Atlas (http://www.ncddc.noaa.gov/website/DataAtlas/atlas.htm) and from the FWC FWRI Marine resources Geographic Information System (MRGIS) Internet Map Server (IMS) (http://ocean.floridamarine.org/mrgis/). Fisheries independent monitoring data are captured in relational data bases in all Gulf States. Fisheries dependent monitoring data (MRIP and the Texas Creel Survey) are generally made available in tabular form.
The majority of the data captured for this project are publicly available and pose no proprietary concerns. However, some data and information from individual research projects and research products generated by the project (Table 6) may be restricted for periods of time to allow publication prior to public release of data. We will develop and adopt a detailed data sharing agreement among project participants during Workshop 1. We do not anticipate that any data sources captured via this project will include or create sensitive information.
The project will create and produce multiple information products, including reports, images, software codes and associated metadata. These products will correspond to the objectives described in **Section c.ii**, and are further detailed in Table 6 with respect to format. All products will be publicly available unless they directly display propriety information described above.

Best practices for consistent data organization and scientific quality assurance, file structure, analysis and data sharing will be followed to ensure data tracking and reproducibility. All existing and produced data will be verified and validated according to data type:

a) Tabular data: checks on file structure (e.g., data delimited by proper columns); checks on file organization and description (e.g., check key parameters, sample identifier, variables); checks on data documentation (e.g., data file name, format, content); review of variables metrics; perform statistical summaries; check locations (e.g., using coordinates to create GIS plots to confirm each location); check for file differences among applications and document changes, if any, in tabular data files (e.g., using ExamDiffPro httt://www.prestosoft.com/edp_examdiffpro.asp).

b) Image vector and raster data: GIS image and vector files information (e.g., data type, files sieze, scale, coordinates, missing data, bands, size of image) will be checked to ensure projection parameters were accurate and to ensure data integrity during network transfer. When standards for definition of processing and quality level of image products are specified, the information will be detailed, following accepted methodology (e.g., https://lpdaac.usgs.gov/lpdaac/products/modis_overview; http://outreach.eos.nasa.gov/EOSDIS_CD-03/docs?proc_levels.htm) (Hook at al. 2010). Additional validation will be performed using a data-check program created in a software package (e.g., Statistical Analysis System, Program R) to search for outliers or erroneous data that were missed in this process. All QA/QC procedures will be carried out using scripted programs so that all steps are documented. Data summaries will be produced after running the data-check program to identify, verify, and correct data prior to analysis. Upon satisfactory verification and validation, data will be certified as passing QA/QC protocols, and marked accordingly.

d.ii Short-term Management: Collection and Processing

All pre-existing data, metadata and project products that are not finalized (i.e. are still being processes or subject to QA/QC) will be compiled and integrated in a Red Drum Data Portal (RDDP). The RDDP will be hosted on replicated institutional servers at the University of Florida with daily back-up. Data integration, interoperability, and management of the RDDP will follow the Hook et al. (2010) guidelines. Consistent data organization practices, file structure and descriptive naming and format will be applied to data sets integrated in the RDDP. Periodic data summaries and project reports will be produced twice a year and shared among project participants, providing another opportunity for QA/QC. Data synthesis and modeling will be carried out using multiple software platforms including a SQL relational database, ArcGIS, and the R statistical package (which will be used for most higher-level analysis and modeling tasks).

The RDDP will be stored and made available via University of Florida Research Computing (UFRC) web servers or GatorBox as appropriate. The public data will be freely accessible via the web servers e.g. at http://bio.rc.ufl.edu/pub/RDDP/ and indexed by search engine crawlers in addition to being listed on the RDDP portal website. Restricted unpublished datasets will be made securely available to authorized users via GatorBox or UFRC web servers e.g. at https://bio.rc.ufl.edu/secure/RDDP/projectN/. Large datasets and downloads by interested parties located on unreliable networks will be managed via secure and public project specific Globus end-points (See http://globus.org for details).

Data management will be the responsibility of Dr. Jynessa Dutka-Gianelli (PostDoc and project data management coordinator) with assistance from the data manager (to be recruited) and under the general oversight of the project director, Dr. Kai Lorenzen. Several other project staff and collaborators will also have important roles in data management.

Data collection: the project will use field data already collected by external agencies and

individuals, including project collaborators. These data will be collated and integrated by the data manager under the guidance of Dr. Dutka-Gianelli and in collaboration with the data collectors (Workshops 1 and 2 will focus on data sharing). Metadata generation: Dr. Dutka-Gianelli (existing data sets) and Dr. Ed Camp (project data outputs and products). Data analysis and modeling: Dr. Ed Camp with specialist input from Dr. Dutka-Gianelli (fisheries data analysis) and Dr. Juliane Struve (geospatial data analysis). Data model and/or database designer: Dr. Dutka-Gianelli and Dr. Ed Camp. External data center or archive: KNB Knowledge Network for Biocomplexity (https://knb.ecoinformatics.org/#). Guidance on data management and database design: University of Florida Research Computing.

d.iii Metadata

Metadata will be comprised of contextual information describing the data in a text based document and CSV format, to provide detailed information of the data, and in an XML standard format, for a more detailed data structure description. Metadata will follow discipline-specific standards (ISO 19115-2 for geospatial data and Ecological Metadata language (EML) for ecological and environmental data) and will be machine-readable. Metadata will include details from pre-existing databases and information associated with data, creating a "data lineage" process to enable sharing of information while identifying dataset origins, and providing a way for data validation and quality control. The metadata entry and management tools will be consistently formatted and reviewed by project personnel prior to datasets utilization.
Additionally, a detailed manual describing the nature of datasets, specification of data, and metadata standards will be compiled. Tutorials will detail the creation of the metadata, data control implementation, and data sharing.

d.iv. Data sharing

Project outputs including data sets captured and synthesized, results of statistical analyses, model codes, and digital products (maps and associated products) will be shared through the Red Drum Data Portal (RDDP) during the lifetime of the project and through the KNB repository during and after the project. Data in both locations will be accessible through web portals and be indexable and discoverable. The key data sets to be captured and synthesized (Table 5) are publicly available and so will be the data products derived from compiling and integrating these data sets. Additional, unpublished data sets supplied by individual researchers and research products generated by the project (Table 6) may be restricted for periods of time to allow publication prior to public release of data. Such restrictions will be observed and the RDDP will contain both publicly available data and restricted data that will securely accessible only by authorized users.
A data sharing policy will be developed by project staff and collaborators during Workshop 1. The policy and all specific agreements with data providers will be upheld at all times. As outlined above, data will be discoverable and accessible through the RDDP and the KNB repository. Any requests for access to restricted data will be handled according to the agreed data sharing policy with input of project personnel and data providers. Data requests will be handled in a timely manner and on a nondiscriminatory basis. All data sets will be accompanied by usage rights statements included in data documentation, so that users of data will be clear what the conditions of use are, and how to acknowledge the data source. Completed products that have been certified via the QA/QC protocols (e.g. tabular data sets, data analyses code, reports, etc.) will be transferred to the open-access, searchable KNB repository to ensure their permanent availability and discoverability. It is anticipated that most, even initially restricted products will eventually be transferred to the KNB repository with full

public access (data from individual research projects typically are restricted for up to three years to allow publication by project researchers). The RDDP will be maintained for at least three years after the end of the project in order to maintain restricted data prior to their transfer to the repository.

d.v Long-term management: curation and accessibility

A primary goal of this project is to make critical information broadly available to other researchers and the public (**Section c.ii, Objective 5**). To accomplish this, all completed products will be moved to a long-term data repository. We plan to use the Knowledge Network for Biocomplexity (KNB) respository (https://knb.ecoinformatics.org/#). KNB contains a large number of related biological and environmental data products including all data generated by the National Center for Ecological Analysis and Synthesis (NCEAS). KNB is easily accessible, searchable, and promotes discoverability of products produced from this project. It accepts all ecological, environmental and geospatial metadata standards specified in this project.
The timing of migration of products from the short-term management (the RDDP server) to the long-term repository will depend on the nature of the product. Publicly available information, as well as any products not intended for publication will be transitioned to the repository as soon as they are certified complete, so that such products will be available in a permanent location as soon as possible. Products containing simulation, derived or compiled data that are considered proprietary by their creators will be migrated to the repository as soon as they are published and/or released from restrictions. In addition to the metadata requirements of the repository, a manual (described in ***d.iii Metadata***) will be made available in the repository, and will be updated to reflect all additions to repository through the life of the project and including five years beyond the project completion.

d.vi Data management budget

UF Research Computing replicated storage (a replica on another disk system in another building is included for the safety of the data) = $ 250 / TB / year (5 years) = $ 1250.

UF Research Computing Data management and database design = $ 50/ hour consultation rate (80 hours) = $4,000

Table 5. Description, type and format of key, existing data. (Table limited to larger, key data sets originating from major monitoring programs)

Data name	Description	Data type (all observational)	Format
Florida Marine Resources GIS	Habitat surveys (seagrass, mangroves, oyster beds, etc.)	Habitat data (environmental)	Geospatial
Northern Gulf of Mexico Ecoregion plan	Surveys of structural habitat	Habitat data (environmental)	Geospatial
Gulf of Mexico Coastal Ocean Observing System	Ocean current observations and predictions	Habitat data (environmental)	Geospatial
Florida Fisheries Independent Monitoring	Abundance indices of young-of-the-year and sub-adult red drum	Fisheries independent data (ecological)	Relational database
Texas Fisheries Independent Sampling	Abundance indices of juvenile, sub-adult and adult red drum	Fisheries independent data (ecological)	Relational Database
Marine Recreational Information Program (MRIP)	Estimates of targeted trips, catch and landings; intercept surveys with angler covariate information	Fisheries dependent data (quant. modeling data)	Tabular data sets
Texas creel survey	Estimates of effort, catch and landings; additional angler covariate information	Fisheries dependent data (quant. modeling data)	Tabular data sets

Table 6. Description of project data output and products to be preserved

Objective	Output name	Output description	Output (type, format)
Obj. 1	Synthesized data sets	Habitat; Fisheries independent; Fisheries dependent	Habitat (derived, geospatial), Fisheries (derived, tabular)
Obj. 2	Hierarchical analyses of spatial recruitment and angler effort	Reports; Instructions for analyses; Data analyses code; Geospatial images	Reports and Instructions (text, PDF/XML); Code (text, .txt); Geospatial (TIFF and GIS)
Obj. 3	Social-ecological regional system model analyses	Reports; Instructions for analyses; Data analyses code;	Reports and Instructions (text, PDF/XML); Code (text, .txt)
Obj. 4	Restoration management strategy evaluation (MSE)	Simulation results; Reports; instruction for analyses; data analyses code;	Simulation (simulated data, CSV); Reports and Instructions (text, PDF/XML); Code (text, .txt)