SOUTHERN CALIFORNIA COASTAL OCEAN OBSERVING SYSTEM (SCCOOS) DATA MANAGEMENT AND COMMUNICATIONS SYSTEM PLAN

1. INTRODUCTION

The U.S. Integrated Ocean Observing System (IOOS) is a nationwide effort to provide access to a wide variety of coastal oceanographic and environmental observations and data. IOOS is a program within the National Ocean Service of the National Oceanic and Atmospheric Administration (NOAA) and a Regional Alliance of the Global Ocean Observing System. In the fall of 2002, the Southern California Ocean Observing System (SCCOOS) began the initial design and development of a stakeholder-driven organizational development strategy to enhance and promote the organization, implementation, and application of a regional coastal ocean observing system in Southern California. The SCCOOS region extends 200 nautical miles offshore of the coastline (the seaward extent of the Exclusive Economic Zone) and includes bays and estuaries. Along with 10 other regional associations (RAs), SCCOOS is designated a Regional Information Coordination Entity (RICE) under the authority of the Integrated Coastal and Ocean Observation System Act of 2009 (ICOOS Act).

As a member of IOOS, SCCOOS has a mandate to collect, organize, and provide access to regional oceanographic data. These data need to be quality reviewed, easily understandable, easily discoverable, electronically accessible, and well organized to allow researchers, policy makers, industry, and the general public to make well-informed decisions. To satisfy this mandate, SCCOOS supports a web-based data portal for the entire region providing ocean, coastal, and relevant watershed environmental data and information products.

The goal of the SCCOOS data management system is to curate multiple data streams from the sensors and models supported by SCCOOS as well as from independent data providers, document the data using IOOS-approved metadata standards, provide these data to users via standard services, and archive the data in appropriate long-term archives. The SCCOOS Data System is based on a service-oriented architecture that employs interoperable systems to enable data discoverability via web services and catalogs. The vision of SCCOOS data management is to be recognized in the ocean observation community as a trusted leader in data quality, interoperability and discoverability.

Effective May 2021, SCCOOS initiated a new partnership with Axiom Data Science (referred to hereafter as Axiom) to provide a standards-based lifecycle data management framework that maximizes the discoverability, accessibility, and usability of data and information products and ensures their sustained use. SCCOOS leverages Axiom's data systems that also support three other Regional Associations (AOOS, CeNCOOS, and SCCOOS) to use common infrastructure which enables the dedication of more funds to system advancements and innovation than

would otherwise be possible. The relationship between SCCOOS and Axiom is a collaborative partnership designed not only to serve the needs of SCCOOS, but also to allow for greater contributions to the larger IOOS community. SCCOOS works closely with Axiom to develop and update data management plans, statements of work, facilitate the flow of data, and ensure a coordinated end to end system. The standards and protocols, and annual work plans of the SCCOOS DMAC Sub-system are revised annually by the SCCOOS DMAC Sub-system Management Team and reviewed by the SCCOOS Data Management Committee. Axiom implements recommended and standard practices as defined by the IOOS Data Management and Communications (DMAC) committee and more specifically those in the Guide for IOOS Data Providers, version 1.0 [2006]. These practices apply to data archive, data discovery, data serving (web-based browsing), data transport (binary access to data), metadata, information technology (IT) security and data quality assurance and quality control (QA/QC). In addition, the system meets SCCOOS data management practices that are guided by regional needs with a focus on quality, interoperability, trust, and discovery.

SCCOOS provides access to catalog-level information and to data sets via web services and the public-facing CalOOS Data Portal, which integrates data from SCCOOS and CeNCOOS as integrated California Ocean Observing System data portal. Both web services and the portal are administered by Axiom. The main web service uses Open-source Project for a Network Data Access Protocol (OPeNDAP), built around the Thematic Real-time Environmental Distributed Data Services (THREDDS) Server (TDS). Axiom maintains an Environmental Research Division's Data Access Program (ERDDAP) server that also provides DAP and web map service (WMS) capabilities. Axiom maintains a Geoserver for geospatial data. Geospatial datasets are accessible via Open Geospatial Consortium (OGC) web feature service (WFS) and WMS services. Some data streams are provided to the Global Telecommunications System (GTS) by providing the data via an intermediary such as National Data Buoy Center (NDBC), which posts the data to the GTS.

SCCOOS maintains a website (https://sccoos.org/) which acts as a higher level interface for data discovery and access. The site provides digestible project descriptions and documentation and serves as a platform to communicate current events and stories of interest for the region. This space allows SCCOOS to bridge science communication with data discovery, by linking directly to data and data products. The SCCOOS web and social media provide our broadest outreach, serving data and information products to a region-wide audience. Both the web and social media provide two-way communication. Users can comment on the information and data products, provide feedback and suggestions, and thus shape the flow of information so that it meets their needs. Our help-desk ticket system tracks and organizes our response to feedback. SCCOOS maintains existing product development and communications activities, internet-based education and outreach, and continues the support for investigators at universities and their

associated outreach activities. SCCOOS produces and implements a communications plan that identifies education and outreach materials to be produced.

For each data stream, dissemination (access nodes) is described by one or more of the following terms: GTS, THREDDS, ERDDAP, WMS, CalOOS Data Portal, SCCOOS website.

SCCOOS is implementing recommended and standard practices as defined by the U.S. Integrated Ocean Observing System (IOOS) Data Management and Communications (DMAC) committee, with specific consideration to meet the core capacity requirements outlined in Contributing Data to IOOS. This will ensure data collected by SCCOOS and member entities is distributed on the SCCOOS web site and is managed according to best practices identified by NOAA/US IOOS. This also ensures that appropriate metadata and QA/QC practices are followed and that the data are of a known quality to the end user. These practices apply to data standards, metadata and data, transport and access, archival, information technology (IT) security, quality control and quality assurance, described in the NOAA IOOS Program Office DMAC White Paper (v1.0), and data management and communications DMAC requirements for IOOS Regional Associations and other IOOS grant recipients who are providing data to IOOS.

The SCCOOS Data Management and Communications System (referred to hereafter as the SCCOOS DMAC Sub-system) must adhere to these practices, and this SCCOOS DMAC Sub-system Plan provides the approach to the necessary implementation, describing how data are ingested, managed and distributed from the source to public dissemination. The SCCOOS DMAC Sub-system Plan is organized as follows:

- Section 2 provides an overview of the SCCOOS Data Management and Communications, describing: the function, goals, and objectives of the SCCOOS DMAC Sub-system and, details related to the SCCOOS DMAC Sub-system Management Team roles.

- Section 3 briefly describes the SCCOOS data resources, defines data categories and asset types, and describes how the data categories are handled in the plan.

- Section 4 presents the SCCOOS DMAC Sub-system statement of work and includes descriptions of the system computing infrastructure including details about the processes related to data ingestion, standards for format and content, metadata and data discovery, quality control procedures (including procedures for data that cannot undergo quality control) and flagging protocols. Additionally, this section covers policies for stewardship, public access and dissemination, data archival and preservation, and data system performance and security measures.

This document, unless superseded, pertains to a period of five years from April 1, 2018 through April 1, 2023.

2. SCCOOS DMAC SUBSYSTEM

The mission of the SCCOOS DMAC Sub-system, is to acquire, archive, and share coastal and marine data and information products to meet the needs of SCCOOS stakeholders and the national US IOOS program. SCCOOS uses a data management system that allows a complex array of oceanographic and environmental data types to be well organized, discoverable, accessible, and understandable. The SCCOOS DMAC Sub-system employs a distributed data management approach, which allows data to seamlessly interchange between participating agencies. The system is composed of an internal master node coupled with external data provider nodes. External data providers include stakeholders, partners, and SCCOOS funded projects who produce, manage, and share data. This distributed configuration increases capacity and technical knowledge within individual groups, allowing them to better meet their own internal data management goals. The distributed architecture leverages hardware, bandwidth, and staff resources across multiple systems and groups. Utilization of currently available external data feeds for sensor, remote sensing, and other data sources improves access to data for SCCOOS users with minimal effort.

Integrating available interoperable data feeds into data access applications and data management systems adds a variety of resources at a low cost. Large quantities of real-time and historical sensor information, remote sensing satellite information, and marine habitat and biological data for the SCCOOS region are openly available for use through interoperability protocols. For example, National Aeronautics and Space Administration (NASA) Earth Observations (NEO) provides an expansive array of long term oceanographic, climate, and atmospheric remote sensing datasets. Real-time and historical sensor data feeds for the SCCOOS region are available for hundreds of sensors via the NDBC, the Center for Operational Oceanographic Products and Services (CO-OPS), National Estuarine Research Reserve System (NERRS) and other NOAA programs. Additional sources of interoperable data include those hosted at NASA's Jet Propulsion Laboratory (JPL), U.S. Geological Survey (USGS) TerraServer, and other research organizations. SCCOOS integrates all of these data and makes them available on the CalOOS Data Portal.

2.1 DMAC SUB-SYSTEM MANAGEMENT GOALS AND OBJECTIVES

The SCCOOS Director, Data and Information Manager, and Product Developer, as well as Axiom comprise the DMAC Sub-system Management Team and are tasked with fulfilling the primary goals and objectives within the SCCOOS Data Management Plan.

Goal 1: Provide Core Data Management Support to the SCCOOS Program

1. Provide technical support for SCCOOS cyberinfrastructure.
2. Develop and maintain web-based data portal.
3. Deliver real-time, delayed-mode and historical data for *in-situ* and remotely-sensed physical, chemical and biological observations.
4. Deliver model-generated outputs, including both nowcasts/forecasts and reanalysis, to SCCOOS users.
5. Implement U.S. IOOS Quality Assurance of Real-Time Oceanographic Data (QARTOD) QA/QC checks for SCCOOS real-time data feeds.
6. Develop and implement processes for archiving into federal archives (e.g. National Centers for Environmental Information (NCEI)).
7. Provides system performance and security measures.

Goal 2: Provide DMAC support to the SCCOOS program

1. Provide overall DMAC project management and oversight.
2. Participate in regional, state, national and international DMAC activities.
3. Engage with data providers to access, understand, and appropriately document data (metadata and QA/QC) that is ingested through the SCCOOS infrastructure.
4. Facilitate communication between different data providers to leverage regional and technical experience.
5. Participate in regional committees and task teams (including teams as determined by the Director, and the joint State-Federal Data Integration Initiative) in order to facilitate data integration and interoperability within the region.

6. Participate in national and cross-regional committees, workshops and task teams in order to further the development of a coordinated approach to IOOS data management.

7. Work closely with the SCCOOS office, other data management awardees if selected, and appropriate advisory committees to implement identified user products, tools and web interfaces; develop product requirements; and beta test and refine products in order to increase their utility.

8. Provide reports as requested.

9. Develop detailed work plans with measurable timelines, deliverables, and performance metrics; and assist with proposal development.

Goal 3: Web Hosting and Support

1. Host and maintain the CalOOS Data Portal at https://data.caloos.org/.

2. Provide access to the user interface and visualization tools, data products, data query and access tools, decision-support tools, agency project tracking systems and databases, as well as IOOS Registry tools.

3. Work with SCCOOS staff, SCCOOS PIs, and member organizations to update the Data Portal periodically, in order to improve access to data, ingest new data, develop new tools, improve clarity and ease of use, and the overall "look and feel."

2.2 DMAC Sub-system Management Team

The SCCOOS DMAC Sub-system Management Team consists of the SCCOOS Director, Program Coordinator, and Data Analyst, and the Axiom Data Team, including the DMAC Sub-system Lead and the Data Management Technical Lead. SCCOOS Staff and subcontractor employee information and CVs are included in Appendix C - DMAC Personnel Resumes. Additional information about management roles and responsibilities, including the advisory committee's role, is provided below.

The SCCOOS DMAC Sub-system is advised by the DMAC committee, a panel of regional experts that advise SCCOOS staff on data management practices and implementation on an ad hoc basis. The DMAC committee is composed of experts from partner institutions who serve on a voluntary basis. The co-investigators on the IOOS award and other grants received by the host

institution have data management responsibilities that are described in the Axiom proposal sub-award statements of work.

2.2.1 Roles and Responsibilities

Director (Clarissa Anderson, SCCOOS). The SCCOOS Director manages operations for RICE certification and standards compliance, of which the SCCOOS DMAC Sub-system is a critical aspect and has direct oversight of the Axiom sub-award effort.

Program Coordinator (Megan Hepner-Medina, SCCOOS). The SCCOOS Program Coordinator is responsible for developing data management plans and statements of work with Axiom, serving as the project manager and as the SCCOOS point of contact for the Axiom Data Team. The Program Coordinator primary responsibilities are to facilitate data ingest, website integration, and general data management coordination. The Program Coordinator works with the Axiom Data Team and data providers to ensure a coordinated system from data ingest through data visualization and dissemination, and integration with the SCCOOS website. The role has direct responsibility for the acquisition, curation, and delivery of SCCOOS data. The role is also responsible for coordinating data management planning and implementation with the IOOS program office and other regional associations, and in this role attends the annual DMAC meeting and other IOOS data management coordination activities (e.g., monthly calls, report and standards review and development). The Program Coordinator liaises with SCCOOS-supported data producers to ensure that appropriate guidelines are followed. The ultimate responsibility for the information management conducted by SCCOOS-supported investigators rests with the SCCOOS Director who implements the sub-awards.

Data Analyst (Data Analyst, SCCOOS). The Product Developer provides information on data sets and assists in reviews of Axiom data services and visualizations. The Product Developer works with the Axiom Data Team on developing and transitioning derived products and advanced visualizations, such as for data collected from multiple depths, from moving platforms, and combinations of data types. The role creates oceanographic products targeted to specific stakeholder needs. The products may be directed by the SCCOOS Information Products Committee. The Product Developer works closely with the Program Manager and Data and Information Manager to formulate and update products. The Product Developer works with national and other regional DMAC teams to harmonize and streamline products.

DMAC Sub-system Lead (Rob Bochenek, Axiom Data Science). The role oversees the Axiom Data Team and the SCCOOS DMAC Sub-system. The DMAC Sub-system Lead together with Axiom's Director of Programs, Stacey Buckelew, contribute to proposal development and general SCCOOS data management reporting requirements. The DMAC Sub-system Lead is the main point of contact for all technical data-related questions and is an expert in managing large

scale datasets related to the SCCOOS mission. The DMAC Sub-system Lead submits semi-annual reports to the SCCOOS Director.

Data Management Technical Lead (Shane St Savage, Axiom Data Science). The role is responsible for implementing IOOS-recommended technologies for the collection, curation, delivery, and archive of SCCOOS data. The technical lead advises SCCOOS on the application of technologies that meet user and stakeholder needs and oversees their implementation.

Data Management Committee. The Data Management committee advises the SCCOOS Director and provides guidance to SCCOOS on both a strategic planning and operational level, regarding the acquisition, curation, and delivery of SCCOOS data. The committee consists of data scientists and informatics experts who are familiar with coastal marine science data, standards and protocols for data stewardship (including IOOS recommended protocols) and familiar with existing and emerging technologies for the browsing and visualization and delivery of scientific data.

2.2.2 Communication

The Axiom Data Team meets with the Program Coordinator weekly to discuss projects status and timelines and communicate any task related issues. Projects involve a kick-off meeting with SCCOOS providing initial guidance, defining project scope and setting project milestones. The Axiom Data Team participates in at least one regional meeting (i.e. State, West Coast Ocean Partnership, or other conference) in addition to the regular IOOS/SCCOOS meetings The Axiom Data Team attends the annual IOOS DMAC meeting, in addition to maintaining communication with IOOS to keep up to date on national IOOS office expectations.

2.2.3 Axiom Data Team

The Axiom Data Team is comprises staff from Axiom Data Science and they are involved with all aspects of the SCCOOS data flow, including data ingestion, creation of metadata, conversion, discovery, maintenance of data feeds, storage, and any necessary archival services. Axiom's primary goal is to gather and serve data important to SCCOOS end users via standard services as recommended by the IOOC and the IOOS Program Office (e.g., OPeNDAP, ERDDAP, THREDDS, etc). It is also tasked with managing and archiving any SCCOOS-funded and non-funded data generated by oceanographic models, buoys, or other devices to enable generation of data products.

The Axiom Data Team is responsible for the design and deployment of a DMAC Sub-system to meet the needs of the SCCOOS user-base. This system must provide the functional components required by IOOS RICE as described in this plan. The team offers comprehensive technical solutions to data management needs, underpinned by a scalable, open source system that uses

existing and emerging software, high performance computer clusters, and interoperability services. This data management system provides an environment that increases the access and use of data by all user groups and allows data management staff to rapidly develop new capabilities and tools to meet emerging user needs.

The Axiom Data Team is dedicated to providing data management and informatics support for SCCOOS, and provides development capabilities for map-based data portals, spatial planning tools, and data management frameworks which transfer and ingest data from external systems via interoperability protocols. The team ensures transparency and communication between client and contractor about design requirements and development progress, and continually researches and employs new technologies to extend the capabilities of digital information and computer analysis systems.

SCCOOS and Axiom personnel maintain regular communication with the U.S. IOOS Program Office through a variety of mechanisms including in-person meetings, phone calls & webinars, email conversations, and GitHub repositories. The continuous communication ensures that the DMAC team and Axiom is aware of all new practices and protocols, as promulgated by the Interagency Ocean Observation Committee (IOOC) and the IOOS Program Office, and understands how and when to implement them.

2.3 SCCOOS Procedures for Evaluating the DMAC Sub-system

The SCCOOS DMAC Sub-system Management Team contracts for professional data management services from Axiom. The Axiom Data Team selection followed a process of broadly soliciting competitive proposals to provide web portal, data management, communication and user product services for SCCOOS for up to 5 years. The Axiom employee CVs have been evaluated and new SCCOOS DMAC Sub-system contributors will be evaluated before starting in such a role. SCCOOS procedures followed during the solicitation, evaluation, and selection of contractual data management support are fully described in SCCOOS Framework for Decision Making. SCCOOS solicited proposals for two sets of services, and proposers were encouraged to bid on one or both of these components, separately or combined.

I. Data Management and Communication Services: Services include providing data management support (data ingestion, metadata, relational database development and maintenance) and communication services (web portal, data clearinghouse, coordination, and communication), building upon the hardware, software, query tools and products developed over the previous years, and following the national IOOS Program guidance.

II. User-Driven Product Development Services: Develop user-driven products and associated interface and visualization tools that will be maintained by and interact with the data system developed under component I above.

The Axiom contract is administered by the SCCOOS Director via the University of California, San Diego (UCSD). The Axiom contract scope of work lists the services which the contractor must be able to provide and this in turn is used to evaluate the quality of work provided by the contractor. The SCCOOS Director coordinates quarterly evaluations of the DMAC services contract including discussion of Axiom DMAC performance. The Axiom subcontract is also reviewed annually as part of the Governing Council program review. Annual reviews take place directly following SCCOOS Fall Science Impact meetings. The SCCOOS Director will use recommendations from the program review to inform future subcontracting agreements.

The Program Coordinator is hired by the SCCOOS Director and reports directly to him/her with annual evaluations. Continued employment of all SCCOOS Program Office staff is dependent upon responsible execution of the duties incumbent to the position they hold and in accordance with UCSD's personnel manual (not publically available).

3. SCCOOS DATA RESOURCES AND ASSET TYPES

The SCCOOS DMAC Sub-system provides data to the public from multiple sources including SCCOOS-funded projects and data from numerous and diverse external federal and non-federal organizations.

3.1 Observational Data Types

The SCCOOS data inventories include multiple types of data, including real-time data, near real-time data, and historical data. SCCOOS defines each data type in a consistent manner with IOOS Guidelines as follows:

- *Real-time data* are ingested, served, and displayed by the SCCOOS DMAC Sub-system at the same frequency the data are collected (and sometimes reported) by the originator with little to no delay. Real-time assets primarily include shore stations, HF Radar, gliders, oceanographic buoys and numerical model data.

- *Near real-time data* are ingested by the SCCOOS DMAC Sub-system at the same frequency that the data are made available; however, there is some delay (hours to days) between data collection and when the data provider makes it available. Examples of near real-time assets include HAB sampling data and derived satellite products.

- *Historical data* are data that are one month old or older. Historical data are sometimes collected in real-time and then archived, and sometimes ingested from local or national archives on request.

3.2 Data Categories

The SCCOOS data types are divided into five major categories that determine the level of documentation and quality control (QC) that is required for the data assets within each category:

1. Federally Sourced Data
2. Model Products
3. Static Data Products
4. Funded Data Streams
5. Regional Partners Data Streams (not SCCOOS funded)

3.2.1 Federally Sourced Data

Federally sourced data incorporated into the CalOOS Data Portal are quality controlled following rigid data management and archival processes by the federal agency collecting the data. These data only require generic documentation by SCCOOS on how these data are ingested and made available to the public (Section 4). As of the writing of this plan, federally sourced data served by the SCCOOS DMAC Sub-system are all exempt from detailed data stream documentation. Federal sources include the National Oceanographic and Atmospheric Administration (NOAA), the U.S. Geological Survey (USGS), the National Weather Service (NWS), and the National Estuarine Research Reserve System (NERRS). See the Appendix B - Federal Data Product Inventory for a complete list.

3.2.2 Model Data Products

Model outputs and products served by the SCCOOS DMAC Sub-system may incorporate or assimilate observational data (e.g., all bathymetric charts served by SCCOOS are from gridded models derived using “true” observations). These models are considered a product that falls outside the realm of “true” observations therefore they are exempt from detailed data stream

documentation. See the Appendix C - Model Data Product Inventory for a complete list.

3.2.3 Static Data Products

SCCOOS static data products are typically derived from observed data, but are displayed in a way that the original data are no longer reproducible and cannot be used to assemble a numerical observational dataset in time or space. Other types of static data products are merely representations of fixed political or legal boundary information. These products fall outside the realm of SCCOOS observations and are exempt from detailed data stream documentation. See the Appendix D - Static Data Product Inventory for a complete list.

3.2.4 Funded Data Streams

Data funded by SCCOOS fall into their own category. The primary processes involved with data management include data ingestion, standards and format, metadata and discovery, quality control, stewardship and preservation, access and dissemination, archival and security. Descriptions of the processes that consistently apply to all data streams are provided in Section 4. Additional data management documentation unique to individual data streams are provided through a systematic Data Stream Plan template that follows the RICE Certification Guidance DMAC requirements (section 1-6) and the NOAA Data Sharing Template. Use of a custom Data Stream Plan template facilitates consistent documentation, and streamlines future additions and edits to existing data stream protocol.

The Data Stream Plans use a consistent and comprehensive set of questions designed to describe how data streams are handled and managed end to end. Grouped parameters may originate from a single platform type (e.g., a mooring that provide temperature, salinity, and dissolved oxygen data, all of which are treated in a standard way); a data type that is handled similarly across all platforms (e.g., webcam imagery); or originate from a single data source (e.g.,Marine Institute at the University of California, Santa Barbara or Scripps Institution of Oceanography (SIO)).

Quality Control descriptions included in the Data Stream Plans may follow one of four paths for a given data stream:

1. Follows prescribed QARTOD guidelines (required for real-time data only if a QARTOD Manual exists for the parameters in the data stream).

2. When QARTOD guidelines do not exist, some other suitable form of QC implementation

is conducted and described;

3. A description of the QC completed by the data provider (e.g., brief description or link to QC protocols performed at the source).

4. Data are considered exempt from QC documentation or requirements if federally sourced.

See the Appendix X - Funded Data Stream Inventory for a complete list including links to individual Data Stream Plans.

3.2.5 Regional Data Streams

Regional Data Streams are defined here as any data resource that does not fit into the exempt categories already discussed: federally sourced data, model product, static data products, or are not funded by SCCOOS. These include, for example, regional data provided by local or state agencies, private companies supporting maritime activities in coastal waters, university projects, and research studies funded and conducted by local entities. None of these data streams served by the SCCOOS DMAC Sub-system originate directly from SCCOOS funding. Most regional data originate from sole source providers affiliated with other entities (research, private, NGO, etc.). Occasionally, a federally sourced data asset is manipulated in some fashion prior to display and, therefore, requires documentation (e.g., federal satellite data that is transformed from a NSIDC-binary format into netCDF). Data streams may be of any data type: real-time, near real-time, historical, or citizen science. Leveraged projects in which SCCOOS helps support but does not fund may also fit into this category. Due to the external nature of these data streams they are usually exempt from detailed data stream documentation. On occasion, however, a data stream that would normally be considered exempt will require documentation in a Data Stream Plan:

- Data products that include representations that can be used to reproduce numerical data in time or space are considered observing data, are treated as a Regional Data Stream and are further documented in a Data Stream Plan.

- A federal data source that is translated or transformed in some way between the source at ingestion to the SCCOOS access point of delivery (e.g., smoothing, block averaging).

See Appendix F - Regional Data Stream Inventory for a complete list including links to individual Data Stream Plans, as applicable.

4. SCCOOS DMAC SUB-SYSTEM ARCHITECTURE AND WORK PLAN

4.1 Computing Cyberinfrastructure

SCCOOS DMAC Sub-system employs a framework for managing a variety of ocean data types (*in-situ* and remotely sensed data streams, multidimensional grids, geographic information system (GIS), and other structured formats). This framework, developed by Axiom, exposes managed data through interoperability systems and uses several user interface tools that allow the data to be discovered and explored by the broader community. Use of this framework to power the SCCOOS DMAC Sub-system enables SCCOOS to rapidly ingest or connect to data sources relevant to SCCOOS and develop advanced user tools and data products efficiently.

The system is divided into four tiers, which separate the suite of technologies composing the system. See Figure 2 for a diagram of the system.

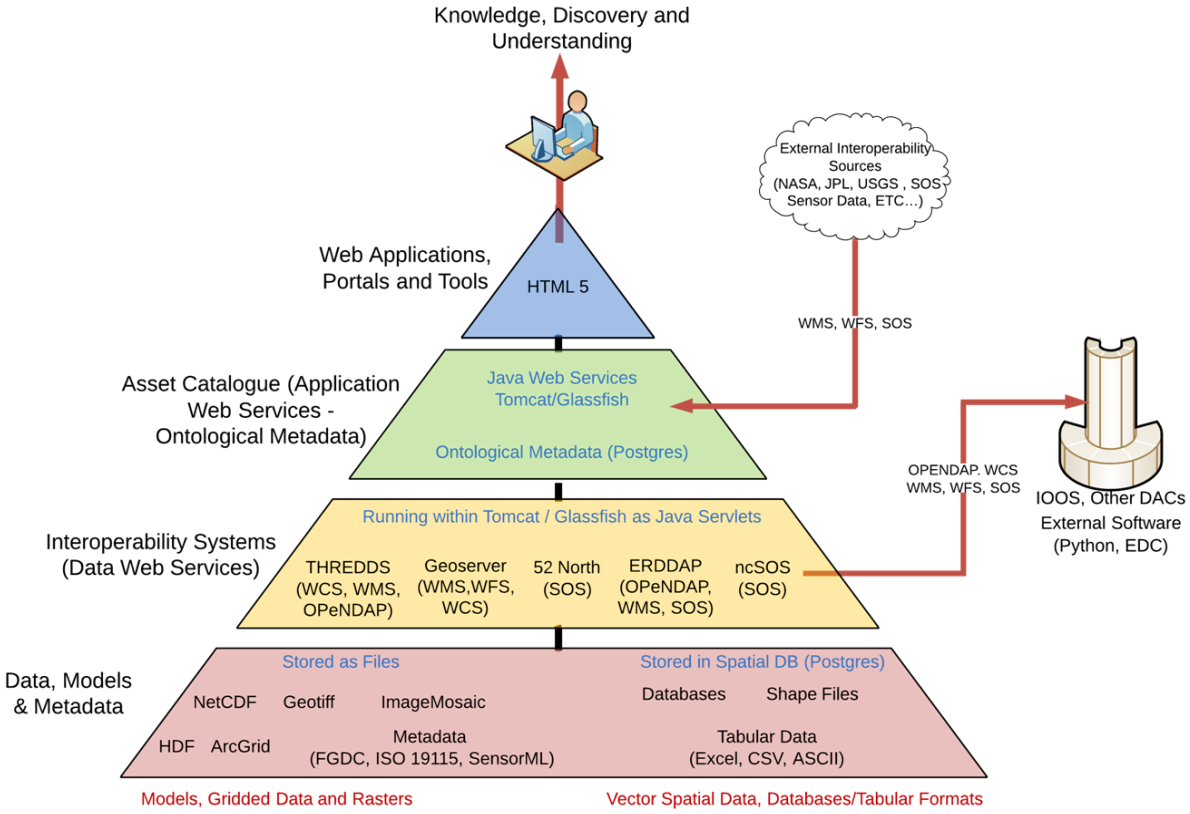

Figure 2. SCCOOS DMAC Sub-system framework showing the flow of data through logical technology tiers, enabling discovery of data that enables understanding the

ocean and coastal environments.

SCCOOS partners' data, models, and metadata are ingested autonomously into the back-end data system through a series of harvesting mechanisms written in Java, Scala, and Python that make use of lower-level interfaces (e.g., file transfer protocol [FTP], hyper text markup language [HTML] and ad hoc service application program interfaces [APIs]). Data files are processed during the ingestion process and loaded into a clustered file storage and database system (GlusterFS and Postgres). A suite of interoperable systems connect to the data storage, including GeoServer, THREDDS, ERDDAP and ncWMS , and they expose data feeds through WFS, WCS, WMS and OPeNDAP protocols. The SCCOOS asset catalog is a database containing ontological information describing the dimensional characteristics (space, time, unit, measured parameter and taxonomy) of each known data resource and how these characteristics relate to each other across data sets. References to both internally- and externally-hosted data feeds are stored in the ontological database and provide the user with a harmonized set of interfaces for consistent access to data and visualizations. Sensors, numerical model output, and remotely sensed observational grids are mapped to common characteristics (space, time, and climate forecast parameter) for comparison across sources. Data sets are further mapped across keywords and, if applicable, Integrated Taxonomic Information System (ITIS) records. The asset catalog also exposes web services providing external access to metadata in the database and provides a method for indexing metadata across multiple formats and types using ElasticSearch, a scalable, Apache Lucene based, clustered search engine. The underlying system architecture works together to allow users to rapidly discover, access and use data through web-based applications and tools developed using modern web development languages and libraries.

4.2 Data Ingestion

Observations and information are ingested into the SCCOOS DMAC Sub-system from a variety of sources, including both historical and real-time observations, forecast, nowcast, and hindcast model outputs, GIS information, and synthesized products that can be useful for layering with other data in the SCCOOS DMAC Sub-system. Each data asset ingested into the system has its own level of data processing maturity and quality with respect to the metadata available.

Data has the ability to be ingested into the system using one of several pathways:

1. Contribution by the originator
2. Direct access or harvest from the originator website or ERDDAP instance (i.e. real-time sensors, delayed mode, or historical data)
3. Auto submission pathway from the Research Workspace

SCCOOS -funded partners provide data to SCCOOS in a timely manner, stipulated in the US IOOS descope proposal. When possible, data are served in real-time. In cases where projects do not produce real-time data, the project PIs are responsible for making sure data becomes accessible by SCCOOS as soon as possible.

4.3 Standards for Format and Content

4.3.1 Shared Data File Formats

SCCOOS provides nearly all data in four open and standardized forms:

1. *Network Common Data Form (NetCDF*) - a self-describing, machine-independent data format that SCCOOS uses primarily for raster (gridded) data. Some data stored as unstructured grids use this format as well.

2. *Comma Separated Values (CSV)* - a human-readable ASCII format that is nearly universally accepted by spreadsheet and programming languages. SCCOOS uses CSV formats to allow users to download: (1) time-series extractions from raster data, and (2) GIS vector and polygon information (e.g., boundaries).

3. *Shapefile* - an open geographic information system format for point, vector, and polygon data. SCCOOS allows users to download shapefiles of static GIS layers such as boundaries, biologic distributions, etc.

4. *Portable Network Graphics (PNG*) - PNG is a lossless image format provided as an alternative to shapefiles in the SCCOOS catalog. PNGs are limited in use as they are pre-projected, pre-scaled, and pre-sized images of data layers. SCCOOS provides PNG files as example WMS requests, which are useful to users who cannot access GIS services and who do not understand how to manipulate WMS requests.

4.3.2 Data Access Points

Access points provide standardized, documented services that allow users to download data from SCCOOS without having to make person-to-person data requests. These standard services are provided through four main platforms:

1. Thematic Realtime Environmental Distributed Data Services (THREDDS) - THREDDS is a set of services that allows for machine and human access to raster data stored in NetCDF formats. THREDDS provides spatial, vertical, and temporal subsetting as well as the

ability to select individual dimension or data variables to reduce file transfer sizes. SCCOOS provides THREDDS access points for raster (gridded) data and discrete time-series observations stored in NetCDF format.

2. Environmental Research Division's Data Access Program (ERDDAP) - ERDDAP is a common data server that provides access to subsetting and downloading data in a variety of formats. SCCOOS provides ERDDAP access to all time-series data in the region, a subset of gridded data, and some GIS-data based products.

3. GeoServer - GeoServer is used to serve image tiles and provide download formats for tabular GIS data.

4. ncWMS - ncWMS is used to serve image tiles for gridded datasets (NetCDF).

Service protocols provided by these platforms include:

1. Open-source Project for a Network Data Access Protocol (OPeNDAP) - OPeNDAP is a protocol that can transfer binary or ASCII data over the web. Like THREDDS, it provides spatial, vertical, and temporal subsetting and the ability to select individual variables to reduce file transfer sizes. Unlike THREDDS, requested data are provided as non-NetCDF, structured output. OPeNDAP output can be imported directly into graphical programs such as GrADS, Ferret, or R. SCCOOS provides OPeNDAP access points for raster and time-series data.

2. Web Map Service (WMS) - WMS provides machine access to images, which can be used by individuals or programs (e.g., tiling services). Accessing programs use GetCapabilities requests to ask for image data in whatever format they require, which allows them to gather image tiles over specific areas with the projections, styles, scales and formats (PNG, JPG, etc.) that fits their needs. SCCOOS provides WMS access points for point, vector, and polygon information, as well as raster data.

3. Web Feature Service (WFS) - this service provides machine access to the vector elements of static layers. SCCOOS provides WFS access points for point, vector, and polygon information, as well as time-series and raster data.

SCCOOS also provides data as downloadable files including NetCDF, CSV, and JSON formats, usually by leveraging services of the platforms described above. Project-specific data are served in their native file formats.

The flow of data from the source to CalOOS Data Portal follows the same general path for all sources as illustrated in the following flow diagram (Figure 3). For cases where the data are

transformed or modified in any way, an explanation is provided in the individual Regional Data Stream Plans for that particular instance. This includes format translations or aggregations of component data streams into an integrated product.

Though SCCOOS relies on local investigators to provide best practices for QA on their activities related to data submitted to SCCOOS, part of the data ingestion process is to establish adequate metadata and provide metadata links that provide the necessary background information to establish the purpose of the data and expected quality.

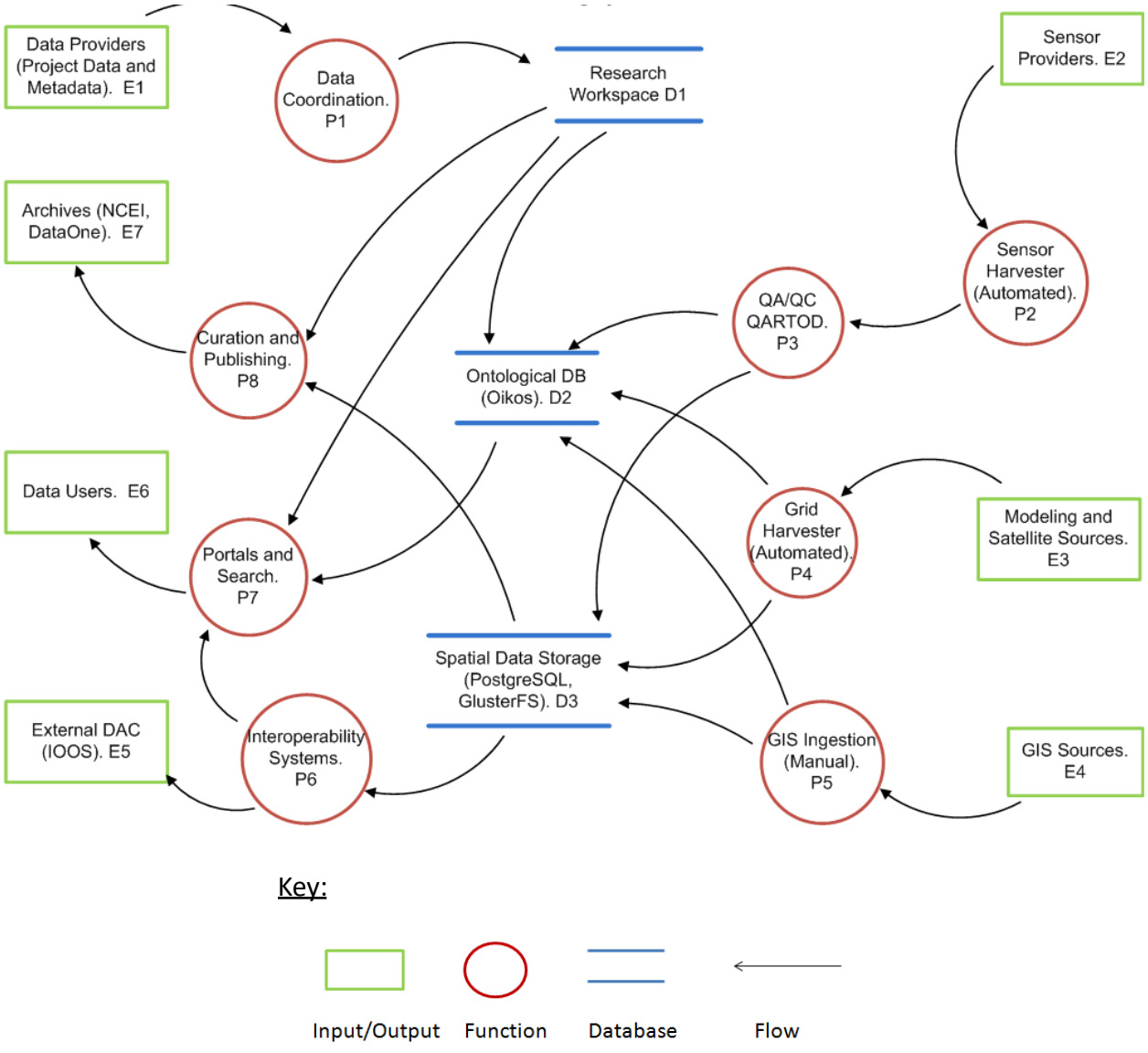

Figure 3. The data flow diagram is a graphical representation of the "flow" of data through the SCCOOS DMAC Sub-system, modeling its process aspects. The diagram makes use of the Yourdon/DeMarco notation. From right to left: the data sources are inputted to the system

(green squares), advance through the system using various processing functions (red circles), are stored within internal databases (blue lines), and are processed and transformed in a standardized manner (red circle) to serve publicly-accessible information products to various outputs (i.e. end users; green squares).

The Data Portal and underpinning infrastructure and service platforms are managed by Axiom Data Science; resources powering these services are leveraged among three other regional associations (AOOS, CeNCOOS, and SCCOOS).

4.4 Metadata and Data Discovery

SCCOOS requires standards-compliant metadata for project-level data (SCCOOS or IOOS-funded projects). Though SCCOOS does not require specific metadata standards for ingesting other types of data, most modern data submittals are accompanied by standard ISO/FGDC metadata records. However, many older data sets come with informal metadata documentation that is variable in terms of completion and detail required by modern standards, and some are only accompanied with narrative information. In these cases, SCCOOS works towards making the source information easily accessible to the end-user by providing links to source data or data providers, and making all available metadata information that came with the data available in the SCCOOS data catalog. Details and availability of metadata are discussed in individual Data Stream Plans.

4.5 Quality Control Procedures

A primary mission of SCCOOS is to serve as a regional data assembly center (DAC), aggregating data from local and federal sources and making them available, accessible, and understandable to the public. Quality assurance (QA) relates to procedures undertaken during the experiment and/or instrument design phases of data collection, ensuring that all the data collected are as accurate and precise as possible. Providing very few data collection devices itself, SCCOOS is reliant on individual data providers to provide adequate QA procedures, and they will not be discussed in this document. SCCOOS Principal Investigators maintain equipment inventories, shipping logs and instrument history logs for equipment owned and/or operated by SCCOOS. All SCCOOS instruments are calibrated, validated, operated, and maintained in accordance with manufacturer's guidance and as recommended by the principal investigators responsible for the equipment.

Quality control (QC) processes implemented by SCCOOS are used to identify and flag or remove bad or suspect data after data collection. Sharing these protocols and quality flags are an

important component of publicly serving data.

4.5.1 SCCOOS Implemented QC Protocols

SCCOOS implements QC policies outlined in the US IOOS Quality Assurance of Real-Time Oceanographic Data (QARTOD) manuals. As new data are ingested into the SCCOOS DMCI DMAC Sub-System, they are assessed and classified accordingly, to allow for full documentation as described in this plan, including Data Stream Plans for new assets that do not come from federal sources and that will be archived by SCCOOS. As new QARTOD protocols are updated and new parameter manuals developed over time, Data Stream Plans will be updated accordingly to include newly required QARTOD implementations. When QARTOD guidelines do not exist for a variable, other suitable form of QC implementation is conducted and described.

Implementation of QARTOD tests by SCCOOS have different processes depending on the data type -- real-time data, historical data, citizen data, and federal data.

Real-time data

The SCCOOS DMAC Sub-System ensures that quality control (QC) standards are implemented and QC flags made available for all real-time data that are not received from a federal source. SCCOOS currently serves various non-federal data streams that require QARTOD QC test implementations. QC procedures differ depending on whether they are implemented by the data provider or SCCOOS.

<u>QC by Data Provider</u>

SCCOOS funded data providers (e.g.,) implement and document the implementation of QARTOD recommended QC tests performed for their data. In these cases, links to these procedures, or a brief summary of the QC performed is provided in the organization's individual Data Stream Plans (see Appendix X). The NWS, NERRS, the IOOS HFR DAC, and the IOOS Glider DAC all ingest and perform extensive QC on the raw data collected by these platforms prior to making them available to the public. In these cases, parameters and/or configuration for the quality test are defined by the data provider. SCCOOS ingests these data and quality flags from the programs for display in the CalOOS Data Portal, and are not required to perform additional QC on these assets. Roll-up summary flags and individual test flags are shown visually in the CalOOS Data Portal with links to the QC documentation made available by the data provider. Flags are also stored alongside the data for download in CSV and netCDF downloads, as well as

via THREDDS and ERDDAP servers. See examples of roll-up summary and individual test flags for a California Polytechnic State University, Center for Coastal Marine Sciences's Morro Bay - BM1 T-Pier station.

QC by SCCOOS

For sources that do not provide quality flags, the SCCOOS DMAC Sub-System runs QARTOD tests after ingesting observation data. Tests are run using the open-source ioos_qc library (https://github.com/ioos/ioos_qc) which implements a suite of QARTOD tests as well as other quality control algorithms. The quality test code and test thresholds are documented and publicly available through the CalOOS Data Portal. Links to the ioos_qc methods used are available both within data charts and on sensor pages within the CalOOS Data Portal. Thresholds used for each test are also viewable on sensor pages and users are linked to the test code in GitHub.

Within one hour after observations are ingested to the SCCOOS DMAC Sub-System, a process is run to calculate flags for the following QARTOD tests, depending on the parameter:

- Gap Test- checks that the times supplied are in monotonically increasing chronological order, and optionally that time intervals between measurements do not exceed a value.
- Syntax Test- checks for parity errors by testing if data can be extracted from the downloaded or scraped data.
- Location Test- checks that a location is within reasonable bounds.
- Gross Range Test- Checks that values are within reasonable range bounds.
- Climatology Test- Checks that values are within reasonable range bounds for a given location and depth
- Spike Test- checks if the difference in values between a data point and its neighbors exceeds a threshold.
- Rate of Change Test- checks if the first order difference of values exceeds a threshold.
- Flat Line Test- checks for consecutively repeated values within a tolerance.

Tests are run for all sensor data that do not already have QC tests applied to it and are applied continuously as new data enter the system. The quality test thresholds can be defined per sensor parameter, when input from the sensor operator or subject matter expert has been obtained. When a specific sensor QC configuration has not been defined it will fall back to a default set of thresholds for each test. For example, the Gross Range thresholds for Air Temp might be (-90C, +60C), and Barometric Pressure might be (800, 1090) mbar.

The quality flag results are made available in the CalOOS Data Portal both visually and for download, similar to source-provided flags described above. By default data is provided with a “rollup” (e.g. summary) flag applied to it. This rollup flag is the worst case of all individual tests (see "Primary Flag" in the QARTOD Data Flags Manual). If any of the flags fail, that data point is not shown in portal visualizations, such as time series plots or anomaly charts, but the data is still available when the dataset is downloaded. An example of the rollup flags for California Polytechnic State University, Center for Coastal Marine Sciences’s Morro Bay - BM1 T-Pier station run by the SCCOOS system can be seen here.

For each individual quality test, the individual test flags are shown visually alongside the data. Within a timeseries chart the “flag statistics” are shown as a stacked bar plot stacked bar plot showing the test results (i.e. pass, fail, suspect, missing data) at each data point alongside the observation data. Users can interact with flags by hovering over them to view a breakdown of individual test results. Additionally, users can turn on and off quality filtered tests entirely or by test results type using a checkbox. The quality flags are responsive to the time binning represented within the chart thereby allowing a user to view summary flags or narrow down to raw data points. An example of the individual test flags for Morro Bay - BM1 T-Pier salinity sensor run by the SCCOOS system can be seen here.

In addition to being viewable, quality test results are available for download in the CalOOS Data Portal. The single rollup flag variable is served alongside the data in CSV downloads, as well as THREDDS and ERDDAP servers. For serving individual quality flags, a second flag variable is also available within the downloaded data for each measured parameter that describes all individual quality flags in one value. Quality tests are described in a standard way, as described by the IOOS Metadata Profile v1.2, in which QARTOD flag variables are associated with data variables using the CF “Ancillary Variables'' approach. An example of the Morro Bay - BM1 T-Pier salinity sensor dataset and embedded quality test results available in ERDDAP is here.

Historical Data

When QARTOD applies, data assets that were previously reporting real-time data and that have had their historical data stored and made available in the SCCOOS DMAC Sub-System, follow the same QC protocols as the real-time data (e.g., non-federal weather data). In these cases, the applicable quality tests are run retroactively for all legacy time series data available within the CalOOS Data Portal. The quality test results are available visually along the time series continuum within the CalOOS Data Portal and for download following the same procedures as the real-time data.

4.6 Stewardship and Preservation

SCCOOS stores ingested data in a secure, professionally managed external facility. Data are stored on storage volume with multiple levels of fault tolerance and can survive multiple server and drive failures. In addition, data are backed up to Backblaze, an offsite backup service provider. SCCOOS currently has total storage space for over 1.8 petabytes of data.
SCCOOS stores all aggregated data, be it real-time sensors, forecasts results, static GIS layers, etc., indefinitely beyond the life of each individual project. This means that real-time sensor feeds will become historical sensor feeds one-month after collection, and it allows SCCOOS to grant users rapid web-based access to all sensor data (federal and nonfederal) since SCCOOS began aggregating feeds. The only assets that are not kept indefinitely in storage are webcam images, NERRS data (as it is strictly prohibited in their terms of service), and forecast products that have been replaced with a more accurate forecast.

4.7 Providing Public Access and Dissemination

All data served on the SCCOOS DMAC Sub-system are fully available to the public and have no data restrictions or embargo periods placed on them. New datasets from either new or current data providers received by SCCOOS are immediately available to the public after data ingestion and documentation is complete; however, they are not added to the searchable data catalog. Datasets are added to the searchable, public catalog only after the data provider is brought into a feedback loop to comment on the metadata, usage notes and citation information regarding their dataset. Once published in the catalog, products are promoted via the SCCOOS website, social media accounts, outreach campaigns, and an email newsletter.

The SCCOOS DMAC Sub-system provides a variety of environmental and socioeconomic data resources in a one stop data portal, free to the public, with data originating from SCCOOS funded data providers, federal and state agencies, local municipalities, academic institutions, research organizations, private companies, non-profit organizations, and community observers. Any data served by the CalOOS Data Portal carries with it the permission to view and access, and carries no privacy or ethical restrictions. Data access is defined here as being permitted to download data through the CalOOS Data Portal. Occasionally, a data sharing agreement between SCCOOS and a data provider will identify the existence of intellectual property rights (IPR) to the data and this is noted in the applicable Data Stream plan. However, IPRs do not restrict access to any of the data that is freely served through the CalOOS Data Portal. IPR information is provided to show its provenance. It is a best practice to always clearly give credit to the data source (the originator) and data provider (in this case SCCOOS) in any work or publications that emanate from using data accessed via the CalOOS Data Portal and to similarly

provide clear data attribution in the CalOOS Data Portal itself.

4.8 Data Archival

As a federally-funded program, SCCOOS is required to submit data it generates to a national archive center. SCCOOS has an active Submission agreement to archive data with the National Center for Environmental Information (NCEI) (see Appendix H - NCEI Archival Agreement). The agreement framework is being renegotiated with NCEI and is in the process of transferring that agreement to Axiom Data Team. The SCCOOS DMAC Sub-system Management Team are working with the NCEI to assist with the transfer and preservation of appropriate data types. SCCOOS has worked and consulted with several NCEI staff members, including Matthew Biddle (former), John Relph, and James Partain, on automating the submission of SCCOOS-funded data assets to the NCEI. Julie Bosch will advise the SCCOOS staff on the data submission forms and all necessary procedures. Automatic SCCOOS submissions for long-term archive at NCEI are currently limited to *in-situ* data from a handful of regional observing networks (listed in Table 1).

SCCOOS also archives data through two national IOOS Data Assembly Centers (DAC): the IOOS HF Radar DAC and the IOOS Glider DAC. Both national DACs are responsible for archiving all of their data, including that submitted by SCCOOS, with NCEI and is beyond the scope of this document. For more information see the National Glider DAC and the National HFR DAC.

Table 1. Funded Data Streams Inventory for Archive of *In-situ* Observations Data to NCEI.

Source	Station Name	External Identifier	NCEI Accession	Status
University of California, Santa Barbara	Stearns Wharf	SCCOOS *in situ* water quality monitoring	0157036	active monthly archive through December 2020
University of California, Irvine	Newport Pier	SCCOOS *in situ* water quality monitoring	0157034	active monthly archive through December 2020
University of California, Los Angeles	Santa Monica Pier	SCCOOS *in situ* water quality monitoring	0157016	active monthly archive through December 2020
Scripps Institution of Oceanography (SIO)	Scripps Pier	SCCOOS *in situ* water quality monitoring	0157035	active monthly archive through December 2020

4.9 Performance and Security

SCCOOS stores observational data on servers located at Scripp's CoLo data center. SCCOOS operates two enterprise servers running Redhat operating system. Each server has Xeon Processors, 96 GB of RAM and shares a JBOD with 12 Terabyte of storage. Servers are backed-up nightly and copies stored offsite at the University of California San Diego Supercomputer Center. Annual disaster and recovery hard drives are sent to the University of California Santa Barbara for additional backup.

Currently, in the event of a catastrophic failure at our primary data center, our WordPress-based SCCOOS and HABMAP websites, including their associated MySQL databases, as well as the SCCOOS ERDDAP and FTP servers, could be quickly recovered with minimal data loss by bringing these servers up in the cloud and restoring the associated data from backup. Computer systems are being migrated to Docker, with associated code in SCCOOS' BitBucket source code repository.

All physical infrastructure required to support the SCCOOS Data Portal and its underpinning systems is located in a data center designed and maintained by Axiom staff. Resources in the data center include more than 2,500 processing cores arranged in a series of interconnected blade arrays, as well as slightly more than 1 petabyte of usable storage that includes multiple redundant backups. Compute nodes and storage nodes are connected over a low latency, converging network fabric (40 Gb/s Infiniband). GlusterFS is employed as a storage software abstraction layer that enables clients and storage servers to exploit data transfer over Remote Direct Memory Access (RDMA) protocols. This configuration enables data throughput from the storage clusters to the compute clusters to reach speeds greater than 160 Gb/s in high-concurrency situations. The SCCOOS DMAC Sub-system provides the following enterprise-level infrastructure capabilities:

- Security and Redundancy: The SCCOOS DMAC Sub-system is maintained by Axiom at its data center, collocated with the Pittock Internet Exchange in Portland, OR, part of the West Coast US internet backbone. There, the data center benefits from the low-latency, high-bandwidth internet connection, and network and power reliability. All data center resources are protected by several levels of onsite redundancy and backup, with offsite backup using Backblaze B2. This design ensures that multiple redundant copies of data exist in addition to web application servers. Several layers of physical hardware (enterprise-level firewalls) and system monitoring software (Nagios) are also in place to provide hardened cyber security.

- Capacity and Performance : High Performance Computing (HPC) has been a component of the SCCOOS DMAC Sub-system technical strategy since early 2011. Axiom operates its own private cloud of compute and storage resources that data managers can provision to specific tasks and roles. The current numbers of processing cores and storage volumes are scalable to allow additional resources to be added as necessary. The large GIS, model, and remote sensing datasets within the system require HPC environments to be visualized and queried over web-based interfaces. Because HPC is achieved through load balancing and parallelization, these types of systems also provide the added bonus of high availability and redundancy.