Towards Automation of Volunteered and Authoritative Bathymetric Data Comparisons

By:

LT Patrick J Debroisse, NOAA

B.S. in Mechanical Engineering, Norwich University (MCV), 2012

THESIS

Submitted to the University of New Hampshire

In Partial Fulfillment of

The Requirements for the Degree of

Master of Science

In

Ocean Engineering: Ocean Mapping

May 2023

This thesis was examined and approved in partial fulfillment of the requirements for the degree of Master of Science in Ocean Engineering: Ocean Mapping by:

Dr. Brian Calder
Associate Director, UNH Center for Coastal and Ocean Mapping
Research Professor, University of New Hampshire

Andrew A. Armstrong, CAPT, NOAA (retired)
Co-Director, NOAA/ UNH Joint Hydrographic Center
Affiliate Professor, University of New Hampshire

Dr. Brian Miles
Senior Research Project Engineer
UNH Center for Coastal and Ocean Mapping

On 19 April 2023

Acknowledgements

The author wishes to acknowledge support from the Office of Coast Survey of the National Ocean Service, National Oceanic and Atmospheric Administration. Coast Survey supported the assignment of the author to the Joint Hydrographic Center/ Center for Coastal and Ocean Mapping (JHC/CCOM) at the University of New Hampshire as a research assistant during the time this research was conducted.

The author would also like to acknowledge his wife Allison and son Gavin for their support throughout these last three years, his thesis advisors Dr. Brian Calder, CAPT Andy Armstrong, NOAA (ret), and Dr. Brian Miles for their guidance, knowledge, and patience. The author would also like to thank the Coast Survey QC Tools Team (Tyanne Faulkes and Giuseppe Masetti) and Eric Younkin for tips and tricks for python programming, as well as the NOAA NBS team (Glen Rice and Katrina Wylie) and the DCDB data manager Georgie Zelenak for their help with research. Finally, the author would like to thank fellow students Clint Marcus, Nick Custer, Igor Vladimirov, Airlie Pickett, Daniel Leite, Adriano Fonseca, Jenna Ehnot for their endless entertainment and for taking time away from their own work to study or otherwise help in many ways. Without these individuals, this work would not have been possible.

The author is a NOAA Corps officer, with salary and personnel support from the NOAA's Office of Marine and Aviation Operations.

The JHC/CCOM is supported under NOAA grant NA20NOS4000196.

Table of Contents

Table of Figures

List of Tables

Abstract

Towards Automation of Volunteered and Authoritative Bathymetric Data Comparisons

By

LT Patrick J. Debroisse, NOAA

University of New Hampshire, May 2023

Hydrographic data, including sonar, positioning, and oceanographic data used for sea floor mapping, are used all over the world for many purposes. Large volumes of high-quality hydrographic data are required to create high-accuracy nautical charts and bathymetric models required by national defense, resource management, shipping, and scientific interest groups, among others. However, across the world, the amount of high-quality data available is limited. Volunteered Bathymetric Information (VBI) is a relatively untapped data source that could be used in many ways such as filling data gaps and informing future data collection expeditions. Determining the quality of VBI, especially Crowdsourced Bathymetry (CSB), has been difficult and time consuming leading to limited use in official nautical charts by national Hydrographic Offices. Despite this, the International Hydrographic Organization continues to collect and store CSB in its Data Centre for Digital Bathymetry (DCDB) in the hopes of widespread future use.

Recent developments in VBI quality assessment have made its regular use more viable, however data discovery, acquisition, and correlation with authoritative data is still a time-consuming and error-prone manual process that must be improved upon before more widespread use of VBI is possible. In the case of CSB, data discovery involves the use of the DCDB web map viewers to identify and request individual files via email.

There is currently no permanent programmatic solution to data discovery in DCDB. In a world of limited staff resources, automating this process will help to increase the speed with which VBI could be assessed for quality and incorporated into nautical charts, bathymetric models, survey planning, and decision-making tools.

Herein, the design of an open source, Python-based tool called VBI Compare is described. VBI Compare automates the data discovery and acquisition phase of VBI quality analysis workflows and allows for VBI data reputation calculations to be initiated. As part of the data discovery process, VBI Compare ensures co-location of VBI and authoritative chart data and displays the data collected and processing status to the user.

Further, the functionality of VBI Compare is demonstrated by a case study in the Houston Ship Channel, United States using the NOAA National Bathymetric Source (NBS) as the source of authoritative data and DCDB CSB data as the VBI input. This case study shows the start-to-finish use of VBI Compare to locate and collect required data and feed it into VBI reputation calculation tools being developed by the Center for Coastal and Ocean Mapping. This case study demonstrates the real-world utility of VBI Compare to a Hydrographic Office for VBI evaluation.

1 Introduction

Hydrographic seafloor mapping data are used all over the world for many purposes including but not limited to safety of navigation, habitat delineation, coastal process monitoring, and geologic studies [1]. Large volumes of high-quality hydrographic data are required to create the high-accuracy models utilized by various interest groups but, on a global scale, the amount of such data available is very limited. This need for better seafloor mapping was recognized by the United States in the Presidential Memorandum on Ocean Mapping, 2019 (84 FR 64699) which calls for a national effort to collect quality hydrographic data for mapping, characterization, security and other purposes in the United States Exclusive Economic Zone (EEZ) [2]. Not only is there a need to fill gaps in existing seafloor mapping coverage, but the sea floor also changes due to natural and anthropogenic processes forcing existing data to be reassessed and for areas to be periodically re-surveyed. As world reliance on oceans and their adjoining waters increases, the need for more reliable data and data sources along with the reevaluation of existing data also increases. In addition to the reliable authoritative data sources currently employed by ocean mapping agencies, potential additional data sources exist with procedures and tools in development to evaluate this data for integration into established pipelines.

1.1 Current State of Ocean Mapping Efforts

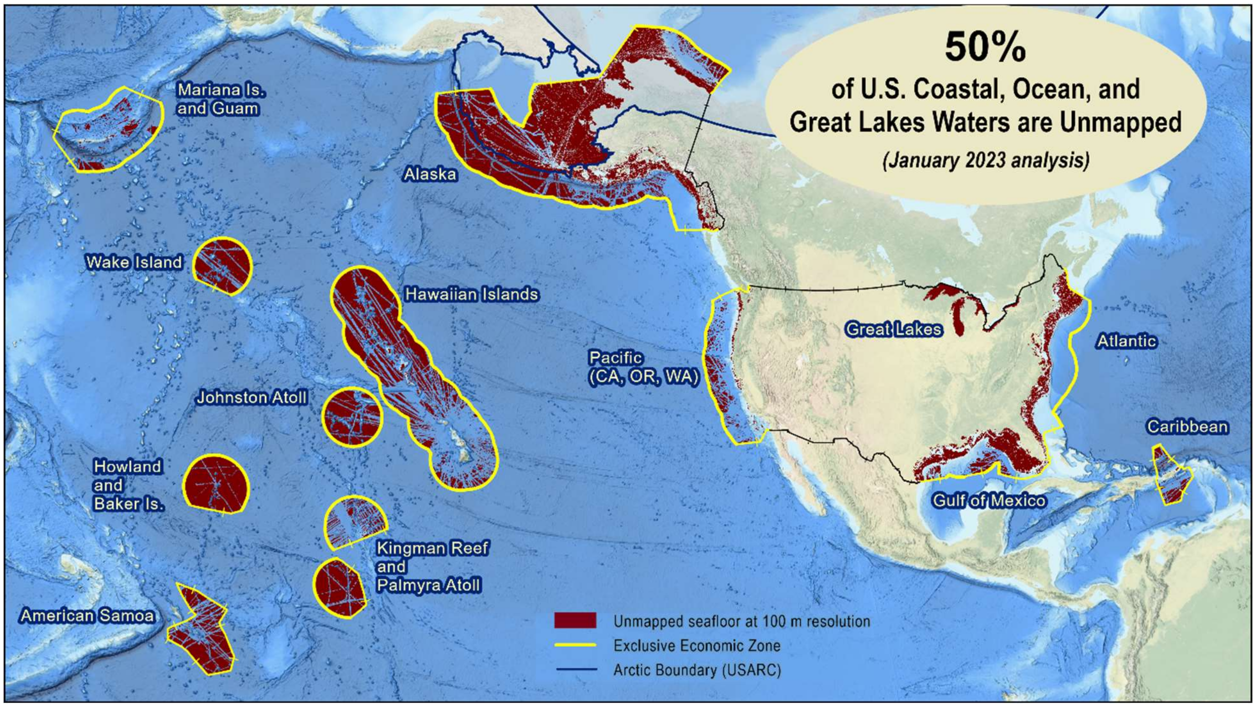


Figure 1: Current area of U.S. waters not mapped to 100m resolution (in red) [3].

The United States has an Exclusive Economic Zone (EEZ) encompassing 3.4 million square nautical miles and at the time of writing (2023), only 50% of this has been mapped to modern standards [2], [3]. Figure 1 shows the current area (red) in the U.S. EEZ still not meeting the 100m resolution standard to be considered mapped as designated by the National Oceanic and Atmospheric Administration (NOAA). Similarly, the Great Lakes Observing System and the Lakebed 2030 Project estimate that only 15% of the Great Lakes are mapped to modern standards (Figure 2) [4].

The Nippon Foundation-GEBCO Seabed 2030 Project has set its own definitions for what is considered mapped based on the depth of the water as shown in Table 1. While there are world charts showing features in the deep ocean such as the Mid-Atlantic Ridge, the resolution of the data used to make those charts may not be fine enough to see smaller oceanic features that are of interest to many users. In shallower water, smaller objects carry greater importance than in deep water. Therefore, data with high vertical and horizontal resolution is required. Achieving similar high resolutions in the deeper parts of the oceans is difficult from the surface, requiring subsea vehicles which are both expensive and slower at surveying compared to surface vessels. The objects of interest in deeper waters tend to be larger than in shallow water, however, allowing for a coarser resolution in these deeper areas for many uses. For these reasons, the resolution requirements to be considered mapped for Seabed 2030 are greater in shallow water than in deep and differ from the NOAA standards.

Table 1: Seabed 2030 Resolution Requirements [5]

Depth range (meters)	Grid cell size(meters)
0- 1,500	100 x 100
1,500- 3,000	200 x 200
3,000- 5,750	400 x 400
5,750- 11,000	800 x 800

On a worldwide scale, according to the Seabed 2030 Project, only an estimated 23.4% of the seafloor has been mapped to these standards (Figure 3) [5]. As these figures demonstrate, large areas of United States EEZ and international waters have no directly measured depth values meeting modern standards. Furthermore, many existing depth values are based on old data collected with low accuracy methods. In this case, while the area may be considered mapped, its quality may be less than desired by a potential user.

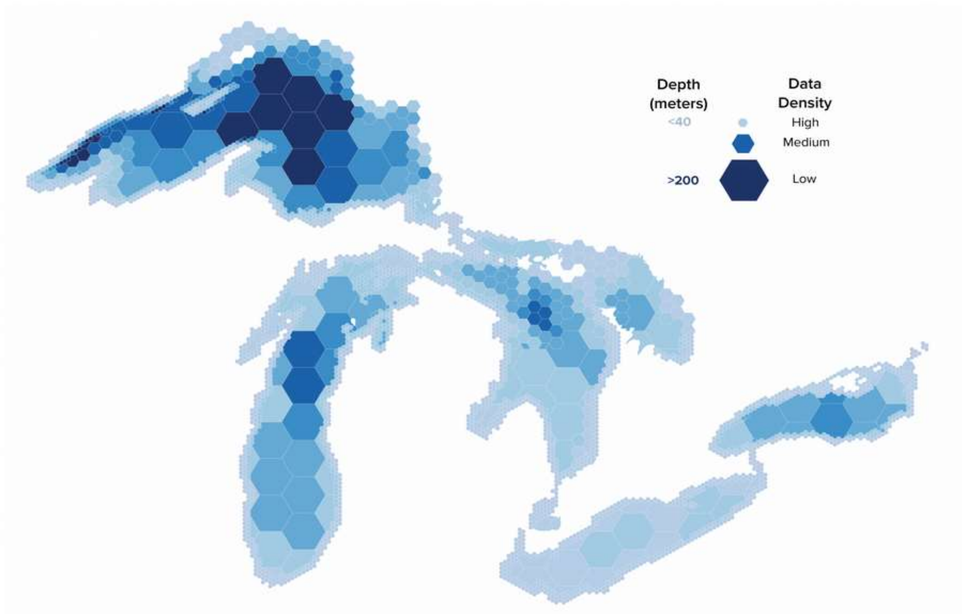


Figure 2: Current data density in the Great lakes. Larger hexagons are less dense [4].

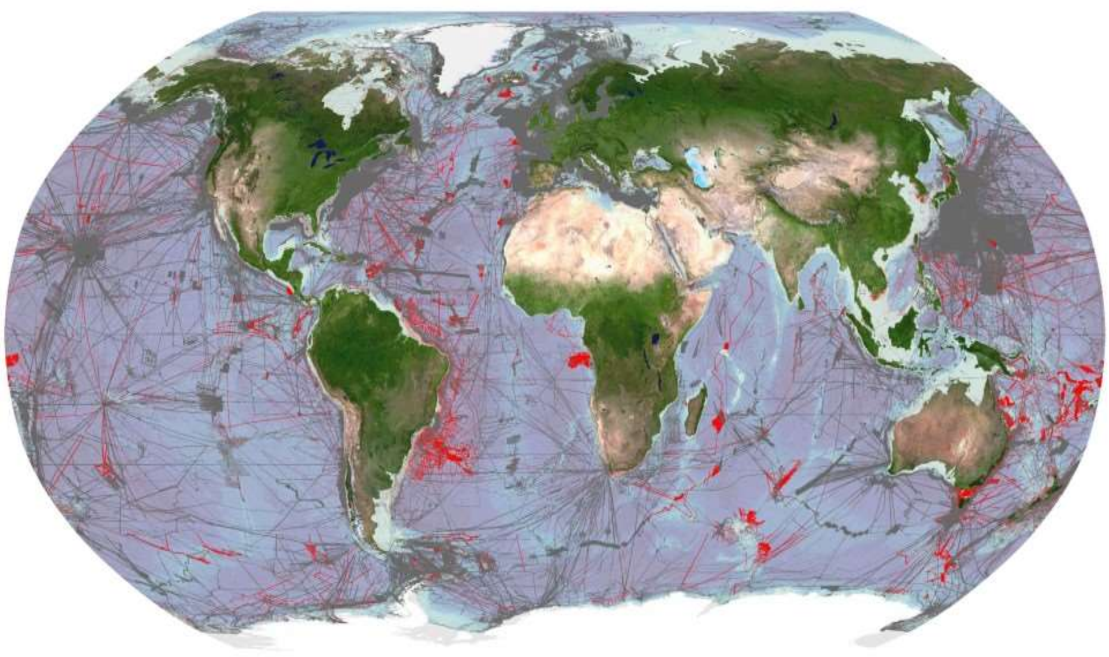


Figure 3: Current area of world oceans not mapped to Seabed 2030 standards (blues/ purples) [5].

1.2 The Need for More Data

The Center for the Blue Economy defines Blue Economy as “the overall contribution of the oceans to economies, the need to address the environmental and ecological sustainability of the oceans, and the ocean economy as a growth opportunity for both developed and developing countries [6].” In the United States, the American Blue Economy contributed $373 billion dollars to the gross domestic product and supported 2.3 million jobs in 2018. In the same year, it outgrew the entire U.S. economy as a whole [7]. Demand for maritime commerce including shipping, tourism, fishing, and more is projected to triple by 2030. Additionally, 40% of the U.S. population lives in coastal counties [7]. As demonstrated above, further growth and development of the Blue Economy will rely on contemporary high-resolution bathymetric data around the world, as shown above.

Current and emerging examples of the need for more hydrographic data abound. Having complete high-resolution data of the EEZ and the entire world oceans will enable many uses such as: understanding ocean phenomena like currents and tides, classifying fisheries resources, monitoring environmental changes, identifying hazards to navigation and safe passage routes, and informing construction and infrastructure projects among other uses [8]. Higher resolution data also allow for more informed decisions and products by policy makers, model producers, design engineers and more. This will have a direct impact on world economies and those who depend on the oceans and connected waters, hence the international desire to map world oceans and connected waterways to modern standards.

In deep water, increased data availability could be used for enhanced geophysical modeling, deep sea infrastructure construction, and vessel safety. The US Navy nuclear powered submarines USS San Francisco in 2005 and the USS Connecticut in 2021, struck uncharted sea mounts while transiting underwater, resulting in loss of life and major damage to the vessels [9], [10]. With finer resolution data in deeper waters, it is possible that these incidents could have been avoided. More data coverage of the world's oceans and lakes can also help support desktop planning surveys for cable and pipe laying. Subsea infrastructure construction requires high resolution data to determine optimal routing and to avoid obstacles that could cause structure damage [11].

In shallower water, delineation of seafloor features can have significant economic repercussions. Shipping companies are building larger vessels, necessarily operating closer to the seafloor to gain economic advantage [12]. By allowing larger ships into ports, vessels can make fewer trips to deliver the same amount of goods and the need for lightering, or partially transferring cargo to another vessel before the ship can enter port, will be reduced saving time and money [13]. Figure 4 explains how just one more inch of allowed draft for a cargo vessel can have a huge economic impact. In the United States, an effort called Precision Marine Navigation [14], along with the future use of IHO S-100 standards [15], [16], is supporting these larger ships operating with critical under-keel clearance by providing high-resolution bathymetry combined with real time weather and tide observations, forecasts, and marine warnings. Increased access to contemporary, high-quality, high-resolution bathymetric data will allow for broader use of precision navigation as bathymetric models are created or updated and higher horizontal and vertical resolutions are obtained.

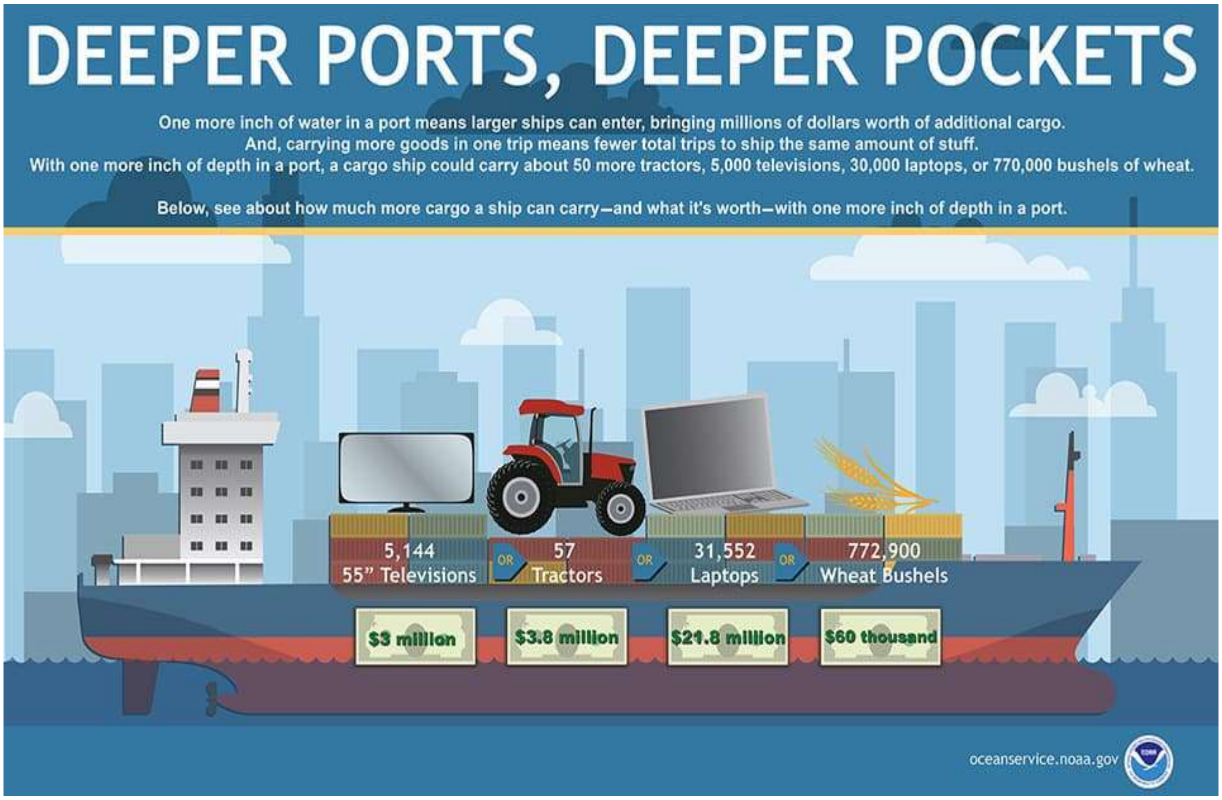


Figure 4: How much is one inch of draft worth? [13]

A wide range of predictive modeling efforts also rely on dense and accurate bathymetric data. For example, storm surge modeling depends on quality bathymetric data to predict set-up of waves as they approach shore. Bathymetry is rarely static in a given area, so these models must be continually updated, especially after a passing storm, to account for sediment transport [17]. With greater access to quality bathymetric data these models will be more accurate and the faster the data is collected, especially after a sediment transport event, the quicker these models can be updated.

Of particular interest is the vertical resolution and total propagated uncertainty of bathymetric data, especially in shallow water. The accuracy of models and charts depends on both horizontal and vertical resolution and uncertainty, but the vertical direction is often given greater weight by Hydrographic Offices (HO) because of the importance of under-keel clearance economically, environmentally, and for vessel safety. Vessel operators tend to steer well clear of known dangers but cannot as easily change their proximity to the seafloor. There are survey systems capable of collecting data with very low vertical uncertainty, but they tend to be prohibitively expensive for most vessels. Other collection methods could provide an opportunity to add or update data in a model or chart especially if high quality bathymetry does not exist or will not be available for a stretch of time.

1.3 Volunteered Bathymetric Information

A relatively untapped source of bathymetric data that can help to fill the gaps described above, or as a tool for ongoing assessment of authoritative chart data, is Volunteered Bathymetric Information (VBI). Volunteered Bathymetric Information is data that has been given to a hydrographic organization from a source other than the normal authoritative sources. These sources could be private industry, academic, recreational vessels, and more. Often, vessels capable of submitting VBI operate in areas that have sparse or aging hydrographic data, or in areas critical for marine traffic. This makes their data desirable to a range of entities such as Hydrographic Offices and forecast modelers, among others.

In the ocean mapping community, the current use of VBI varies by Hydrographic Office. Data is often collected into databases such as the IHO's Data Centre for Digital Bathymetry (DCDB) Crowdsourced Bathymetry (CSB) server for storage and dissemination and is increasingly being used by HOs as a source of monitoring for bathymetric change. The Canadian Hydrographic Service has also used VBI to update some nautical charts directly, but this is still a rare occurrence due to high uncertainties in raw VBI data and the time expense to process the data into a product that meets IHO or national standards [18].

Large commercial vessels are required to carry certified echo-sounders and Global Navigation Satellite System (GNSS) equipment according to the International Maritime Organization Safety of Life at Sea (SOLAS) [19]. Many non-commercial craft are also fitted with this equipment for safety of navigation purposes. Data from these types of systems are what is called Crowdsourced Bathymetry by the IHO and can be voluntarily submitted to the DCDB CSB database. In many cases, a vessel would only need to add or enable a data logger to collect, format, and transmit the required data [20]. Some safety of navigation systems are already fitted with a means of properly packaging and submitting these data with little burden on the user. This means all vessels subject to SOLAS, and many that are not, are likely capable of collecting and supplying CSB with minimal effort. Even a small fraction of these vessels could constitute a large force multiplier when it comes to mapping the world oceans.

The nature of volunteered data varies widely, from sources capable of achieving International Standards (IHO S-44) [21] to data collected with equipment and vessels not originally intended for mapping purposes. Common sources of the data housed in the DCDB CSB database range from personal recreation craft such as sailboats and yachts, to commercial ships, to purpose-built scientific vessels. Minimally, the CSB submitted to DCDB must include depth from an echo sounder, GNSS locations and associated time stamps while the sonar is running, and the time of each sonar sounding

[20]. The DCDB ignores the quality of the sounding data stored in the CSB repository. It does not necessarily have motion sensor, tide, or sound speed data common with high grade mapping data. Further, CSB can be filtered or censored at the request of the nation-state in whose waters the data are collected based on their individual legal agreements with the IHO or because the nation-state does not have an agreement with the IHO to support CSB [20]. Additionally, corrections for tidal variations and equipment installation offsets may be absent from the metadata that accompany CSB data. CSB is just one example of VBI data, but there is a lot of interest in using it because of the potential volume that could be available, and the large area covered by vessels capable of collecting it.

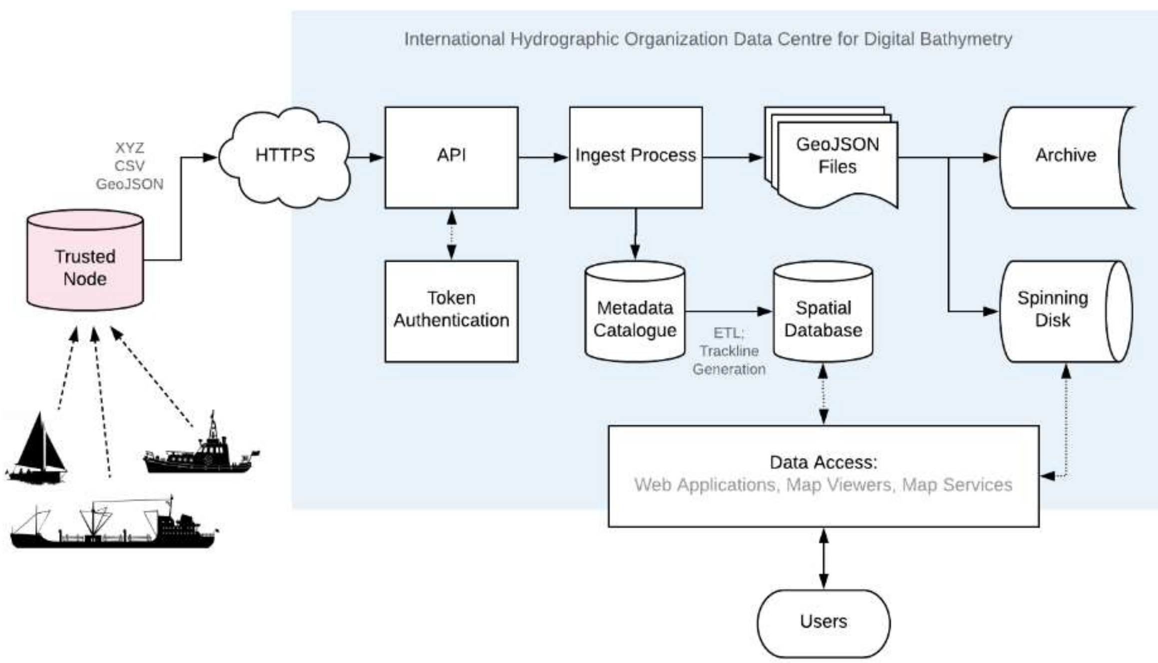


Figure 5: VBI data flow from vessel to user [20].

After collection by a vessel, CSB can be contributed to the DCDB, which is hosted by NOAA at the National Centers for Environmental Information (NCEI). Unless restricted by the contributor, data that is supplied to the DCDB is freely available to the public via a web interface, giving it potential for a wide range of uses. Data collected by the vessel is brought to the DCDB via trusted nodes. Trusted nodes are a network of organizations that liaise between vessels and the DCDB to provide technical support and to ensure data meet the standards set by IHO guidance on Crowdsourced Bathymetry publication B-12, thus ensuring a standard product [20]. The path of CSB data from vessel to public interface is shown in Figure 5.

Far more vessels capable of collecting VBI transit the oceans than those capable of authoritative surveying, making VBI a potentially valuable data stream. Raw VBI may not necessarily contain the stringent uncertainty statistics and metadata required of authoritative sources however, making it difficult to use for authoritative mapping purposes. The CSB housed in the DCDB CSB database is not necessarily expected to have uncertainty values as low as authoritative data because most of the equipment is not purpose built nor installed for survey operations. This does not mean the data is useless. It is possible that some VBI sources could attain a low uncertainty, depending upon available metadata, installation methods, and equipment. Processing this data to determine uncertainty is possible and is key to determining its value to an HO. Processed VBI could constitute a data source with extensive benefit, and depending upon the area, VBI could be the only option available making it a critical data source for that area.

In areas where authoritative hydrographic surveys are lacking or non-existent, quality VBI, even with higher uncertainty, could be used to fill gaps, verify derived bathymetry [22], and assess existing surveys. Adding VBI to the arsenal of an HO could help to avoid the allisions experienced by the Connecticut or San Francisco or aid in the planning of safer and shorter undersea cable runs.

Similarly, VBI with lower-uncertainty could provide an additional data source for hydrographic models in shallow areas. Perhaps more importantly, trustworthy VBI could be used by an HO to monitor these critical areas for bathymetric change. Vessels capable of collecting VBI transit some of these areas far more often than authoritative surveys are conducted. Data from these vessels could provide early warning to an HO about changes in the local bathymetry. This could help influence asset allocation for an HO and safety decisions in a port [23]. For the same reasons, VBI data could be a vital asset to predictive models by filling gaps and confirming or updating existing data.

1.4 Authoritative Bathymetric Data Collection Methods

While other sources such as LIDAR or satellite systems are sometimes used, the bulk of high-resolution low-uncertainty data collection by hydrographic agencies around the globe is conducted by vessels using sonar technology. Often these vessels and systems are owned and operated, or contracted, by the state hydrographic service itself. These data are held to high accuracy standards, meeting or exceeding the International Hydrographic Organization's (IHO) S-44, Standards for Hydrographic Surveys [21]. Generally, vessels collecting data that can meet these standards are equipped with high-end sonar systems including the sonar itself, position and attitude sensors, and processing computers. Unlike VBI data, authoritative data has correctors applied to the soundings during processing to include vessel offsets, tidal corrections, and sound speed data. These systems are expensive as are the vessels they are deployed on and require extensive training of the surveyor. This fact means that the world fleet of hydrographic ships (including government, military, academic, and private industry) is relatively small and specialized and those whose data are directly accepted by hydrographic agencies is even smaller.

Multibeam sonar systems are used by mapping vessels to cover wide swaths of the ocean floor. Overlapping these swaths can create a full coverage three-dimensional model of the sea floor. These types of surveys typically have high density and low-uncertainty which makes them particularly useful for high-accuracy models. Side scan sonars are also common in the mapping fleet for sea floor imaging. They are popular for object searches over large areas because they are capable of even wider swaths than multibeam systems in shallow water. Side scan sonars do not collect depth data, generally, so they are often augmented with single beam or multibeam data as well. Phase measuring bathymetric side scan (PMBS) systems are another method of collecting hydrographic data. PMBS systems provide a wide swath width like a standard side scan but add the ability to measure depth. These characteristics make them particularly useful in shallow water and in surveys intended to search for marine debris [24].

Collecting authoritative data is slow because of the lack of available vessels and the stringent standards the data must meet. The shallower the water, the more this is amplified as the standards become more rigorous and the physics of the sonar systems cause inefficiencies. The result of this effort, however, is a highly accurate archive model of the seafloor. Since the speed of a survey cannot be increased easily, integrating with other quality datasets could reduce the extent of areas needing to be surveyed and increase the monitoring of previously surveyed areas allowing for better allocation of hydrographic resources.

Processing VBI data to determine its value towards this goal remains a challenge. However, in 2021 Calder developed and tested a method of determining the reputation score of VBI observers, and the uncertainty of their data, based on the degree to which their data concur with collocated authoritative data, providing an indication of data quality [25]. Case study data were located in Puget Sound near Seattle, Washington. This method of ranking VBI observers and estimating the uncertainty of their data could help HOs integrate these data sets into their charting and survey planning processes. For instance, data from a vessel with a high reputation and low uncertainty may be trusted enough to be charted directly while those with lesser reputations or higher uncertainty might need further corroboration before charting or asset allocation for resurvey. Additionally, this data could be used to evaluate current products. For example, multiple highly reputed observers consistently disagreeing with charted data in a given geographic area may indicate the need for new surveys [25].

1.5 How Programmatic Discovery of Data will Accelerate Research on and use of Volunteered Bathymetric Information.

In the following sections, a data discovery and acquisition tool called VBI Compare is described. VBI Compare is able to automate the discovery of CSB data from the IHO DCDB CSB database and an authoritative source, feeding these data sets into the VBI observer reputation calculation algorithm developed by Calder [25].

The purpose in this work is to eliminate the need for manual data discovery via web interfaces and allow for reputation calculations to be performed on data sets from any geographic location rapidly. The initial intended use of VBI Compare is to retrieve and stage datasets to test, refine, and verify the Calder algorithm in additional geographic locations more easily. Beyond that, VBI Compare is designed to facilitate accurate data discovery based on user needs and produce meaningful and repeatable data quality assessments for vessels in any study area, allowing VBI data to be more rapidly and widely applied to hydrographic modelling. Here, a case study is described using NOAA's National Bathymetric Source authoritative data and the DCDB CSB database to show how VBI Compare may be adopted by HOs to handle VBI in their areas of responsibility. This case study uses the DCDB Crowdsourced data as the VBI source because it is a substantial international resource and available for free from a public database. Other forms of VBI exist and the Calder algorithm can complete comparisons with that data, as well. The accessibility of the CSB data and the growing interest in it makes CSB the prime VBI data source for initial testing and eventual release of these reputation calculation tools.

1.6 Motivation

To calculate VBI observer reputation and data uncertainty using the Calder method, two data types are required: VBI and authoritative data. CSB, as discussed previously, is stored by the IHO DCDB and is publicly available. The source of authoritative data depends on the needs of the HO employing the tool depending upon where and how their trusted data is stored. It could be their own database, or some external database containing data trusted to be the best available for the area it covers.

In this case study, data from the NOAA National Bathymetric Source (NBS) [26] is used as it is the authoritative public data source for the United States Office of Coast Survey, in the areas where it has been compiled.

In addition to the data requirements, a VBI reputation calculation requires a file management or access scheme that allows it to retrieve and store the required data. Local and cloud processing are both options for the execution of a data quality calculation but have different input requirements to be executed. Additionally, file management is critical to ensure the most up to date data available is used in the calculation to ensure the greatest possible accuracy.

Correlation of the data discovered from each database and placed into a file management structure is the final pre-requisite for reputation calculations. Similar to file management, different methods are required depending upon whether the algorithm is deployed locally or in the cloud.

Currently, these three steps are time-consuming individual manual processes prone to error or incompleteness. A programmatic solution could make this workflow less cumbersome.

2 Program Architecture and Design

2.1 Architecture

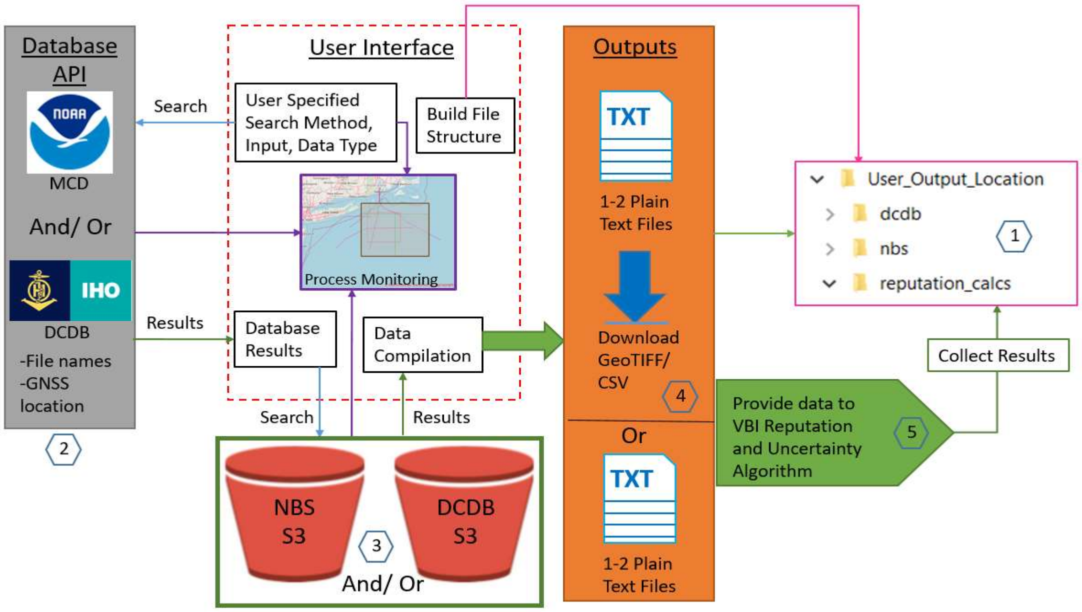


Figure 6: VBI Compare architecture.

VBI Compare is a desktop program produced using the open-source Python language. It provides a single interface for data discovery across the two databases it interacts with: DCDB for CSB and the authoritative database of choice. VBI Compare consists of five major steps as shown in Figure 6. Step one (1) of the program builds a folder structure for data handling. Data discovery and acquisition comprises in step two (2), determining which authoritative or VBI data the user is looking for, and step three, (3) cross references those results with the files available in the Amazon Web Service (AWS) Simple Storage Service (S3) [27] bucket for the respective data type. In step four (4), collected data files are either downloaded locally or the URLs to each file are written into plain text files (allowing for cloud processing without necessarily downloading the files), depending on user needs. VBI Compare then initiates the Calder VBI reputation and uncertainty algorithm in step five (5) if certain constraints are met. Each of these steps are further explained below.

VBI Compare was developed as a desktop app because many existing hydrographic tool sets are already packaged in desktop based platforms such as The Center for Coastal and Ocean Mapping (CCOM)/ NOAA Hydroffice [28] or the NOAA/ CCOM Pydro suite [29]. Importantly, creating this program as a desktop application will allow others to make edits to the program locally for their own needs without having to go to the trouble of setting up and securing web hosting as would be required for a web-based tool. For instance, an HO may want to replace the use of the NBS as the authoritative source with a source of their own.

Using open-source tools like Python and QGIS (which is used to provide the GIS functionality embedded into VBI Compare) will allow users to deploy or update the program to their specific needs without the need to pay licensing fees. Additionally, Python is the chosen programming language of CCOM and the NOAA Hydrographic Systems and Technology Branch (HSTB), which will help VBI Compare to be adopted alongside or integrated with software suites like Hydroffice and Pydro. These suites are publicly available sets of hydrographic quality control, management, and documentation tools which makes them a good fit for future use of this program.

VBI Compare consists of an input window graphic user interface (GUI) where the user is asked for necessary inputs required for data discovery and file handling. The discovery tool collects user inputs to query the databases described above. The input window allows for search criteria to be customized by the user to launch the VBI observer reputation calculation algorithm for the area, vessel, or chart of interest, and initializes file management by requiring input of a user-defined output folder. A monitoring sub-window allows for visualization of the data discovery processing status and the data collected. On the monitoring window, the map display is the primary data visualization tool showing how many files were collected and their geographic location. It allows the user to confirm that the data collected covers the area of interest and is collocated with other data collected. Additionally, text outputs and progress bars allow the user to monitor and confirm progress. This simple interface was designed to be easily used by operators with varying experience levels.

2.2 Data Discovery

To complete a VBI data comparison, the minimum set of VBI data required includes location, depth, and the observation times of these locations and depths. Additional metadata may also be desired to help increase accuracy including vessel offsets or sound speed profiles, for example. In the case of the Calder reputation algorithm, vessel length is an additional required input used to determine if a vessel is moving by determining if the ship has moved more than its length of a certain amount of time. For the purposes of this design, the data within the DCDB and the NBS meet the specific needs for VBI reputation calculation, but they are housed in divergent databases each with its own data storage, access, and metadata.

2.2.1 IHO Data Centre for Digital Bathymetry

The DCDB CSB database houses about 1.5 million linear nautical miles of VBI data from all over the world (as of 2022). These data sets have been standardized for metadata and format by trusted nodes prior to submission to the database. The DCDB has two databases that contain the same CSB. Each contains one file per vessel track line saved in differing formats. The first contains GeoJSON [30] files and is accessed through either a web map interface (Figure 7) or the DCDB CSB (MapServer) ArcGIS REST Service application programming interface (API) [31] using manual input. The second database is an AWS S3 bucket, or cloud data container, which holds Comma Separated Value (CSV) [32] files (Figure 8). The data in the S3 bucket is arranged in directories by the year, month, and day the data was received by the DCDB. Both data types contain depths, time, and location along with other metadata, meeting the VBI needs of the comparison algorithm.

The GeoJSON server does not allow for direct download of the files it contains. When using the web interface, the user must supply an email address to send a File Transfer Protocol (FTP) link for local file downloads. The API query only results in a JSON output rather than downloadable data or links to the data on the server. On the other hand, the S3 bucket does allow for file downloads or collection of URL addresses directly to the data from the bucket itself. Currently, however there is no direct way to query the S3 bucket to determine the required data.

Because the S3 bucket contains the required data and access to it is relatively easy through downloads or direct URL call, this is the database of choice for obtaining VBI. Using the S3 bucket also allows for the possibility of deployment of the reputation algorithm to the cloud. To overcome the data search issue, the REST API is used programmatically as the search mechanism. The file names in the GeoJSON server and the S3 bucket are similar, so using the API to determine the desired files and translating those names to the S3 bucket naming convention to access them is the method used.

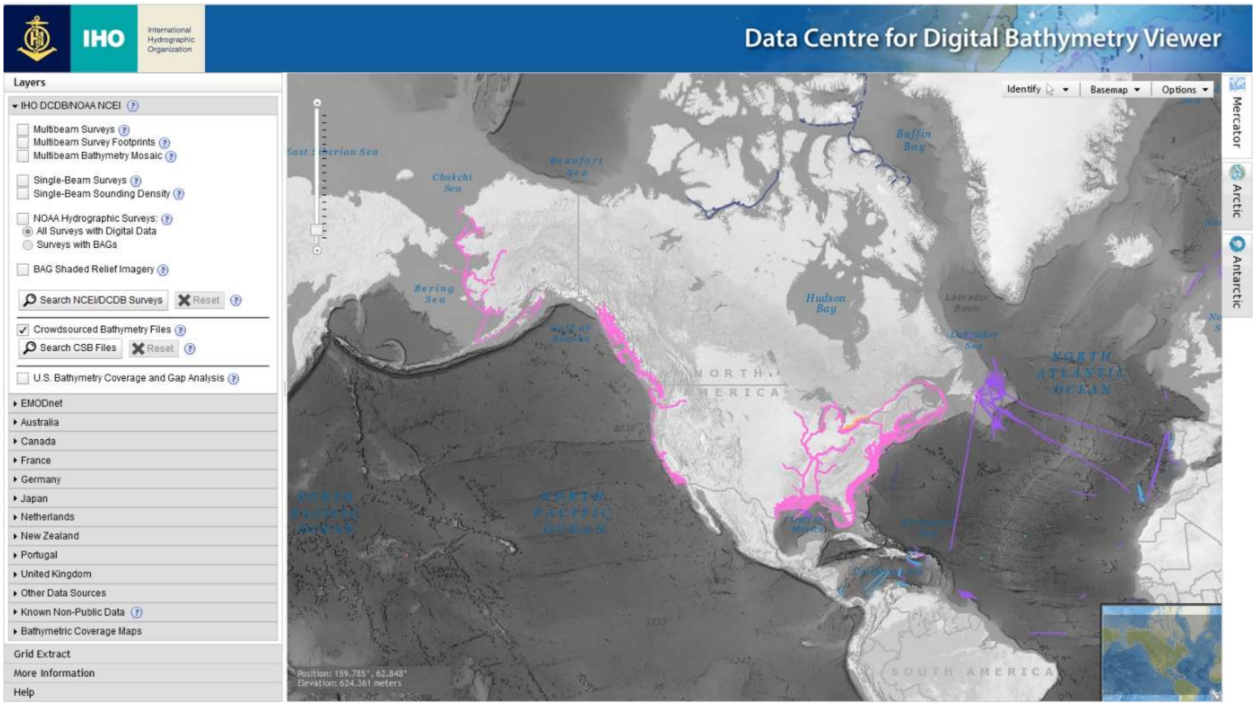


Figure 7: DCDB Web Interface with VBI data displayed [47]

UNIQUE_ID	FILE_UUID	LON	LAT	DEPTH	TIME	PLATFORM_NAME	PROVIDER
PGS-83e03090-0999-11eb-b6ea-98be942a5b5a	17037f997727cd7a40f89446e36ddce6	46.860967	-47.890945	139	2021-08-06T11:00:02.000Z	Ramform Atlas	PGS
PGS-83e03090-0999-11eb-b6ea-98be942a5b5a	17037f997727cd7a40f89446e36ddce6	46.860964	-47.8907	139.1	2021-08-06T11:00:09.000Z	Ramform Atlas	PGS
PGS-83e03090-0999-11eb-b6ea-98be942a5b5a	17037f997727cd7a40f89446e36ddce6	46.860961	-47.890454	139.1	2021-08-06T11:00:15.000Z	Ramform Atlas	PGS
PGS-83e03090-0999-11eb-b6ea-98be942a5b5a	17037f997727cd7a40f89446e36ddce6	46.860958	-47.890206	139	2021-08-06T11:00:21.000Z	Ramform Atlas	PGS
PGS-83e03090-0999-11eb-b6ea-98be942a5b5a	17037f997727cd7a40f89446e36ddce6	46.860958	-47.890206	139	2021-08-06T11:00:27.000Z	Ramform Atlas	PGS
PGS-83e03090-0999-11eb-b6ea-98be942a5b5a	17037f997727cd7a40f89446e36ddce6	46.860954	-47.889964	139.2	2021-08-06T11:00:33.000Z	Ramform Atlas	PGS
PGS-83e03090-0999-11eb-b6ea-98be942a5b5a	17037f997727cd7a40f89446e36ddce6	46.860951	-47.889715	138.9	2021-08-06T11:00:39.000Z	Ramform Atlas	PGS
PGS-83e03090-0999-11eb-b6ea-98be942a5b5a	17037f997727cd7a40f89446e36ddce6	46.860949	-47.889471	138.8	2021-08-06T11:00:45.000Z	Ramform Atlas	PGS
PGS-83e03090-0999-11eb-b6ea-98be942a5b5a	17037f997727cd7a40f89446e36ddce6	46.860947	-47.889226	138.8	2021-08-06T11:00:51.000Z	Ramform Atlas	PGS

Figure 8: Example DCDB CSV file for VBI data

2.2.2 NOAA National Bathymetric Source

VBI Compare uses the NBS S3 Bucket which houses authoritative data in the GeoTIFF format [33]. The bucket contains a folder for each NBS tile area which contains the GeoTIFF and an associated xml file. The GeoTIFF files in the NBS bucket are called

BlueTopo files [26]. BlueTopo is a new data type produced by NOAA's Office of Coast Survey containing the best available bathymetric data for a given chart area they cover. BlueTopo files are three- layer GeoTIFFs with elevation, uncertainty, and raster attribute tables, meeting the minimum requirement of location and depth for VBI comparison. These files contain data from various contributors for the given area and could contain preliminary data still being processed through the NOAA pipeline. The data is also on the North American Vertical Datum of 1988 (NAVD88) orthometric datum which is the official nation-wide vertical datum of the United States [34]. Nautical Chart vertical datum in the United States is Mean Lower Low Water (MLLW) which is a more localized datum influenced by the tide but relatable to NAVD 88 mathematically [35]. These preliminary data and datum factors make BlueTopo files not suitable for navigation [36]. The Calder reputation algorithm accounts for the datum difference by applying tidal corrections and datum offsets as part of the reputation calculation process [25].

NBS uses its own tile convention, separate from that of the NOAA Marine Chart Division (MCD) electronic chart titles created for use in navigation equipment. MCD is the nautical chart production agency of the United States. MCD uses a tile set at specific dimensions and resolutions that correspond to the areas and characteristics of their Electronic Navigation Charts (ENCs). NBS uses a different tile scheme with a wider range of resolution options. While many of the NBS tiles do line up with chart tiles geographically, they may be at a different, often finer, resolution. This allows the NBS to have data tiles at greater resolutions than those created by MCD granting the user the finest resolution of data possible in each area where NBS is compiled. The downside is that it makes searching for a specific tile based on a navigational chart difficult because they use different naming conventions.

Currently, the NBS does not have a web map interface or API server query capability. NBS data is accessed directly in the S3 bucket and can be searched through the use of a GeoPackage [37] file (Figure 9) housed at the top level of the database. This file can be queried to yield the desired file names and URLs to the data directly in the bucket.

The lack of a web map interface and thus ability to visualize an NBS query without third party software is addressed using the geographic information system (GIS) monitoring window of VBI Compare as explained later. This allows the user to visualize the data retrieved by VBI Compare and confirm collocation with the area or data of interest.

As with the DCDB S3 bucket, easy access to the data in the cloud makes this the database of choice for authoritative data. To overcome the naming convention issue between ENC and NBS tiles, geographic area queries are executed against the NBS GeoPackage file. The area used is obtained from the ENC charts of interest using a programmatic API query of the MCD ENC Rescheme Status database [38]. Alternatively, the GeoPackage file can be queried using a user defined area. While the NBS database only covers areas where NOAA has charting authority, it shows how an authoritative database controlled by an HO can be used to compare to the DCDB CSB data to determine reputation and uncertainty.

Figure 9: NBS tile scheme as of March 2023

2.2.3 Database Interoperation

Collection of each individual data type from its respective database is possible using VBI Compare, but to execute a reputation calculation both data sets are required. Further, for a calculation to be completed both data sets must be collocated. This necessitates a query of one database using the results of the other. While these databases and files are divergent as described previously, they both utilize the World Geodetic System of 1984 (WGS 84) [39] for positioning both within the data files themselves and the query interfaces, making geographic queries possible across databases.

In VBI Compare, regardless of how the first database is searched, whether by specific chart or vessel or by geographic area, the second database is queried using the geographic results of the first. In the case where the DCDB is searched first, track line geometry is collected and the NBS GeoPackage is then searched using an intersection query against the track line geometry. If the NBS is searched first, then tile outline geometry is collected and the DCDB API is searched using an intersection query with the tile geometry as input.

2.3 Data Discovery Use Cases

Table 2 outlines several potential research questions that might prompt the use of VBI Compare and the use cases that may address each question. There are sixteen different use cases that VBI Compare can execute as seen in Table 3. Eight of these use cases (1- 8) result in data collection for only one data type while the other eight (9-16) will result in a collection of both. Reputation calculations can only be performed on queries that result in both data types. The use cases that result in only one data type remain available to the user as data collection options for other needs and for testing purposes. Workflows that result in queries of both data sets finish with writing a Windows batch file that can run the Calder algorithm code in addition to the collected files from both data sets. Having these use cases available gives the user options to use the program in multiple ways based on individual needs.

Table 2: VBI Compare Use Case Decision Matrix

	Scientific Question	VBI Compare Use Cases
1	What data is available for a given chart, vessel, or area?	1-8
2	What is the reputation of the data for a specific chart(s)?	9-10
3	What is the reputation of all charts intersecting an area of interest?	11-12
4	What is the reputation of a specific vessel(s)?	13-14
5	What is the reputation for all vessels crossing a specific area?	15-16

Table 3: Sixteen Use Cases of VBI Compare

	Primary Data	Get Secondary Data?	Input Type	Output Type	Comparison Possible?
1	NBS	No	Chart List	Local Download	No
2	NBS	No	Chart List	URL List	No
3	NBS	No	Geographic Area	Local Download	No
4	NBS	No	Geographic Area	URL List	No
5	DCDB	No	Vessel List	Local Download	No
6	DCDB	No	Vessel List	URL List	No
7	DCDB	No	Geographic Area	Local Download	No
8	DCDB	No	Geographic Area	URL List	No
9	NBS	Yes	Chart List	Local Download	Yes
10	NBS	Yes	Chart List	URL List	Yes
11	NBS	Yes	Geographic Area	Local Download	Yes
12	NBS	Yes	Geographic Area	URL List	Yes
13	DCDB	Yes	Vessel List	Local Download	Yes
14	DCDB	Yes	Vessel List	URL List	Yes
15	DCDB	Yes	Geographic Area	Local Download	Yes
16	DCDB	Yes	Geographic Area	URL List	Yes

2.4 VBI Compare Data Discovery Operation

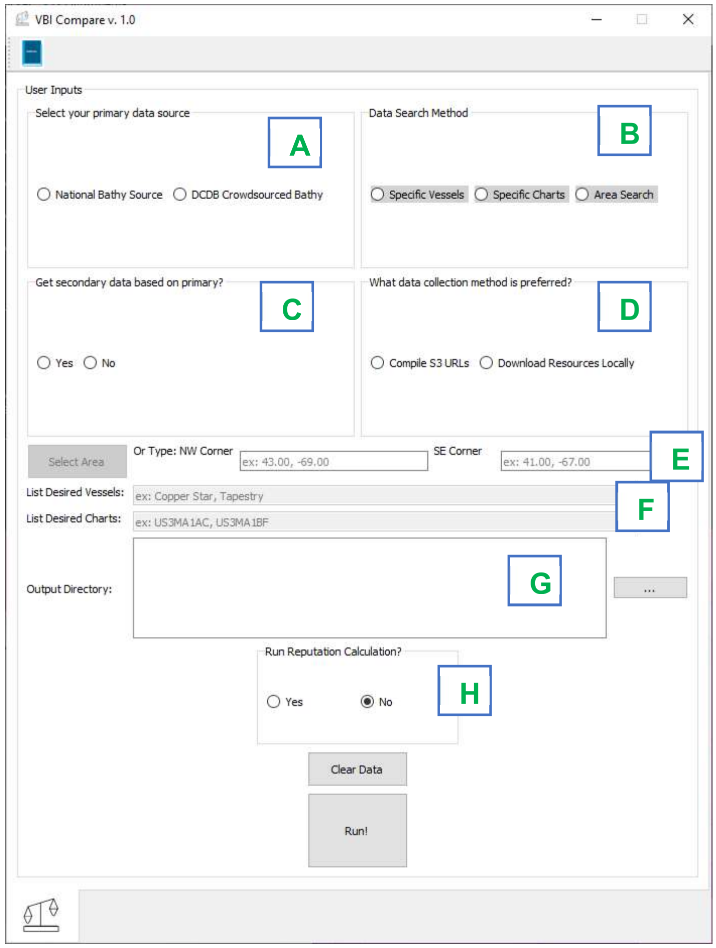


Figure 10: VBI Compare Input Window

VBI Compare requires several inputs to execute the data discovery process described above, which are collected using the input window (Figure 10). The first input [A] is the primary data source. If a user has a specific set of charts or vessels in mind for comparison, this is where that data type of interest is specified. If the user is searching by area without any specific data type in mind, either VBI or NBS can be selected for the primary source. This selection drives the initial data discovery and acquisition process VBI Compare performs. The second requirement [B] is a search method, either a specific vessel, chart, or geographic area. This is the criteria used to search the primary database. The tools to supply the specific search data [E and F] become available depending upon the search method selected. Finally, the user must determine and indicate if they are interested in the secondary data type or not [C]. If they are, both databases will be searched and the results from the primary database search will be used to inform that of the second by providing the input search geometry for the second database query. If not, then only the selected primary database will be searched. This makes it important for the user to determine which data set is driving their comparison needs.

The user constrains the results of their data query by providing either a geographic search area or specific vessel(s) or chart(s) depending on their chosen primary data source. This is done via the input window (Figure 10 (E and F)) by either typing the area, chart number, or vessel name. In the case of area, the user may also use the “Select Area” tool in the GUI (Figure 11).

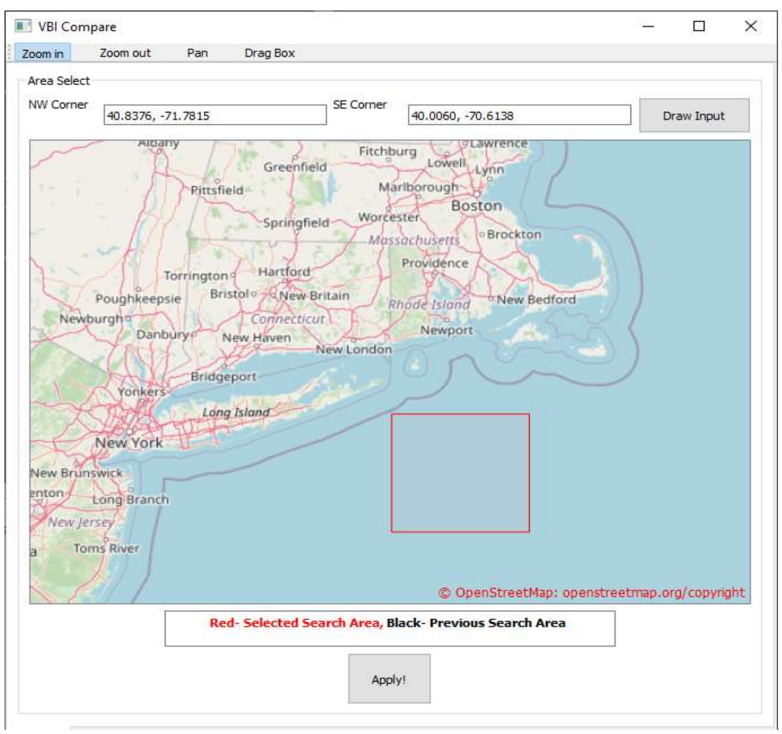


Figure 11: VBI Compare Area Select Tool

2.5 Data Collection and File Management

Once data are discovered in their respective server locations, a method of collection and management of that data for reputation calculation purposes becomes necessary. There are two options for deployment of the calculation algorithm: local and cloud. Each has its own necessary inputs so the script can find the intended data. Currently, file management is a manual process of creating directories and ensuring the proper files populate those directories, which is prone to error.

The Calder reputation calculation algorithm can take in the CSV (or GeoJSON) files from the DCDB and the GeoTIFF files from NBS for processing. Access to those files requires a local download or internet connectivity during analysis to access data in the cloud. A programmatically constructed directory tree-based structure allows predictable file management capable of ensuring the desired files are supplied to the calculation algorithm each time. Two directory structures were considered in developing VBI Compare. The first was to collect all the files discovered by each query into one central directory built per run of VBI Compare. This option would allow the developer to constrain the data passed to the calculation algorithm to this central directory ensuring no unnecessary files were retained. The downside to this method is that VBI Compare would download files each run regardless of whether or not the files were already downloaded on a previous run, increasing the storage requirements of the local machine. Alternatively, the second option was to construct a directory tree which houses the files corresponding to a specific tile or vessel in a folder to themselves along with index files detailing their locations. Doing so allows the program to determine if a file was previously downloaded or if new data is available that supersedes the current local copy. In the case that newer files are to be downloaded, old files could be moved to archives. The index files are used to direct the calculation algorithm to where it can access the discovered data and allow the user to edit which data is used during a reputation calculation without directly affecting any data files. This method reduces the local storage requirement while still achieving the necessary file management and thus was chosen as the file management method for local downloads. There are no size limits to local downloads, so attention to detail is required by the user to ensure their data request matches their needs without additional extraneous area.

An alternative to local processing is cloud deployment of the algorithm. To make this method most beneficial, a file collection and management scheme for data residing in the cloud is required. Rather than download the data, cloud processing can be done through accessing the data directly in the database it resides. Avoiding the need to download files locally makes this method faster and may lead to cost savings because data egress fees are reduced. The Calder calculation algorithm does not need to read all values of the authoritative data, only those cells that contain data that are coincident to VBI observations, so the entire GeoTIFF file is usually not needed. To direct the algorithm to the data, it is provided the URL to the files in its respective database. There are two URL options to access the file: HTTP and S3. While both function well, the S3 method is faster for the Calder algorithm, so this is the URL type which VBI Compare collects.

Once the files are downloaded or the S3 URLs collected, they must be packaged for input to the algorithm. Rather than pass each file individually, a plain text file containing the local address or URL of the files is generated. Using one text file for NBS and one for DCDB data per calculation allows VBI Compare to direct the algorithm to the specific files the user intends while excluding those they do not.

2.5.1 VBI Compare Directory Structure

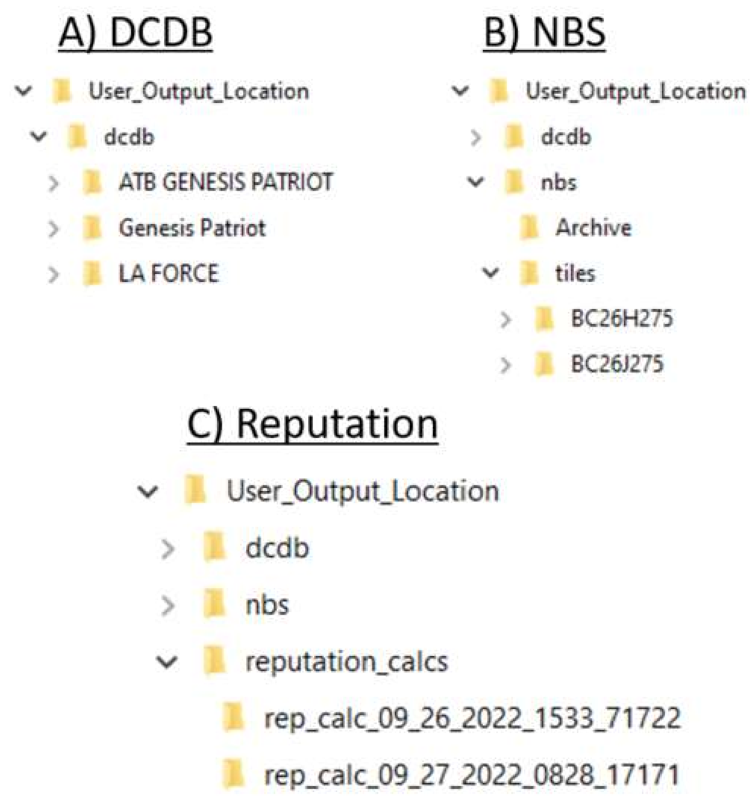


Figure 12: Example File Structures

VBI Compare programmatically builds a directory structure (Figure 12) on the local machine to manage the files as described above. The user is required to supply an output folder in the input window (Figure 10 (G)). This is the folder to which the file hierarchy is built and where data will be stored. The hierarchy separates the various data from VBI Compare into three subfolders, one for each data type: [A] DCDB; [B] authoritative (NBS in this case); and [C] the files specifically created for reputation calculation. Within these three folders, further subfolders are made to organize the data.

In the DCDB folder [A], a subfolder will be created for each vessel. Any locally downloaded CSV files will be downloaded to these vessel folders. In the NBS folder [B], a subfolder is created for each tile. Any locally downloaded NBS GeoTIFF files will be saved to these tile-named folders. In the reputation calculation folder [C], a subfolder for each run of the program is created. VBI Compare will place the batch file created in each run into these subfolders. Additionally, the files created by the VBI observer reputation and data uncertainty calculator will also be saved in these folders alongside their respective batch file.

In the case of S3 URL compilation, there is no need for specific vessel or tile folders since they are not downloaded. Instead, the text files are saved in the base folder (NBS or DCDB) for their respective data types. This is also the location where the plain text files containing local file addresses will be saved when conducting local processing. The NBS GeoPackage is also saved to the NBS folder for data discovery purposes.

Figure 13 shows an example VBI Compare workflow for an area search of both data types with VBI as the primary source. It includes the steps for data management and discovery.

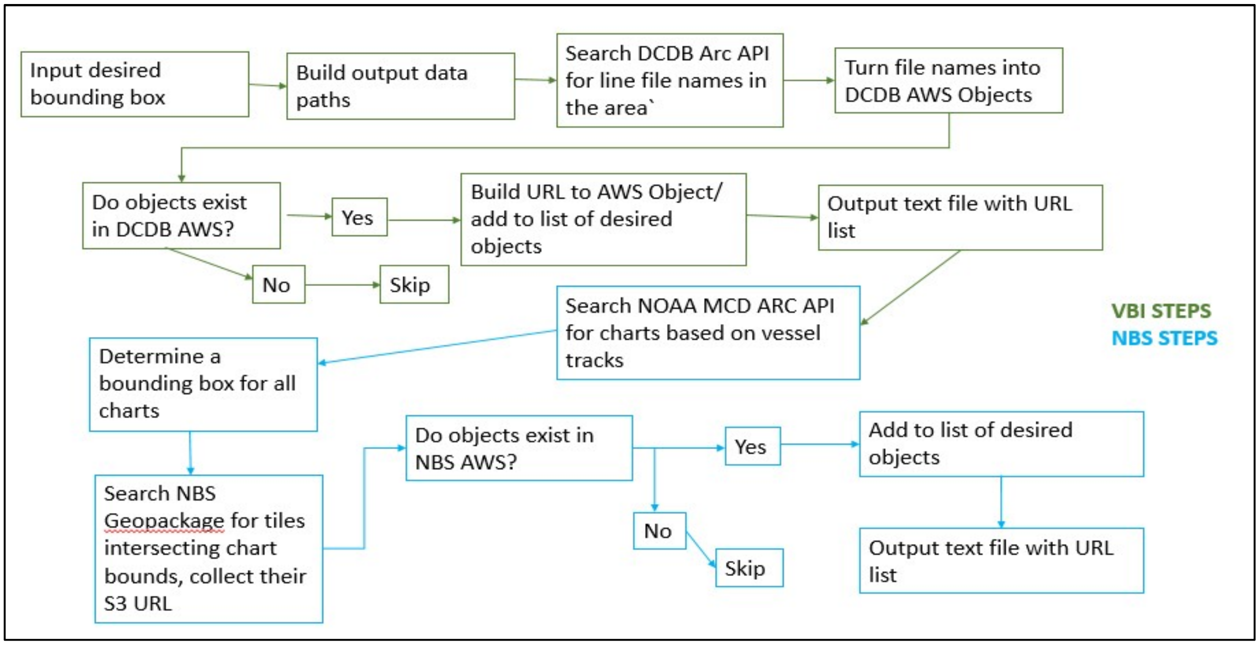


Figure 13: Example VBI Compare Workflow.

2.6 Integration with the VBI Observer Reputation and Data Uncertainty Algorithm

The final step of using VBI Compare is to integrate data discovered from the DCDB and NBS with the VBI reputation and data uncertainty calculation algorithm. VBI Compare creates a Windows batch file (.bat) containing the series of commands needed to load VBI and authoritative data and conduct the VBI observer reputation and uncertainty calculations. VBI Compare can also be easily updated to support Unix shell script creation.

The user is shown the current status of the reputation calculation via a progress bar and text update on the VBI Compare monitoring window. VBI Compare gives the user the option to run the batch file immediately, or to save it for later. If saved for later, the user would simply run the batch file from its directory by double clicking on it or running it from a command prompt.

3 Graphic User Interface

A graphic user interface is required to gather the inputs and parameters from the user and subsequently display critical information back to them. All the various GUI windows for VBI Compare were constructed using Qt and QGIS for Python. VBI Compare contains one input window as described in section 2.4, and one monitoring window described below. Additionally, a sub-window was created for visual identification of a geographic search area if an area search is desired as shown in Figure 11. There are additional message windows and warnings that are generated as well for error handling purposes (Figure 14).

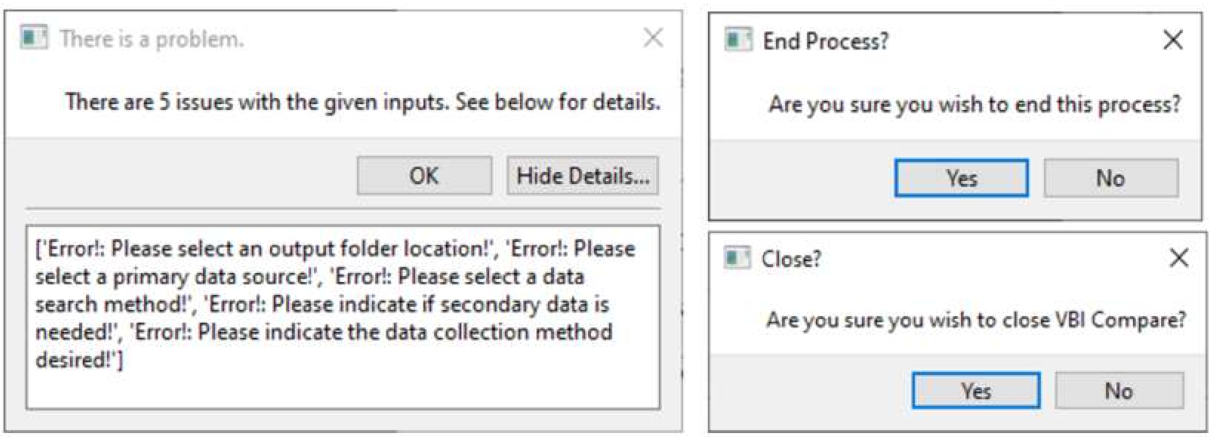


Figure 14: Example warning windows.

Qt was chosen because it is a popular cross platform graphical user interface framework that is well supported in Python and employed in other programs built by NOAA. The need for data visualization necessitated the use of GIS display. QGIS is built using Qt making it easily integrated into the VBI Compare GUI.

3.1 Input Validation and Error Handling

Manual data searches can be error prone as the inputs to a search are not necessarily validated prior to execution. Section 2.4 described the minimum search criteria required to ensure that the data of interest is acquired by VBI Compare. Any missing or incorrect data could result in a query failure or erroneous results.

To limit the chances of misapplying an input for one data search type to another, VBI Compare leverages Qt's ability to selectively enable the functionality of certain input options based on the inputs to others. Additionally, data validation is performed after the user clicks "Run!" but before data discovery is executed. This automatic validation ensures that proper formatting is followed and that all required inputs are supplied. The format of geographic coordinates and data directory names are vital to this process and validation ensures that these conditions are met. Several warning windows, like those shown in Figure 14, will display prompting the user to fix any errors discovered during validation. For example, if the user does not supply an output directory, VBI Compare will raise an error and the program will not run until the errors are fixed.

Even with validated inputs, warnings and errors are still possible during the interaction with the databases. Warnings typically do not terminate data discovery but often indicate some issue encountered during the data search. Errors, however, usually result in a query failure. Typically, these issues are not caused by any user input. Instead, they are programmatic responses to some internal issue with the database itself. Warnings and errors of this nature will be printed in the monitoring window as data is accessed and collected. Examples of database warnings or errors may include when a desired file is not found, if the program cannot access a required server, or if files are being moved to an archive. This built-in error handling is intended to reduce the chances of implementation errors leading to lost time and erroneous results. Printed warnings on the monitor serve only to keep the user informed as to the status of the program in executing their request.

3.2 Visualization

Another objective of VBI Compare is the visualization of the data being collected for validation. The web interface for the DCDB allows the user to select an area or line and visualize it on a GIS window, but the NBS does not currently have this capability. In either case, the user cannot see both data sets at the same time in the same GIS viewer. VBI compare allows this type of visualization as part of the monitoring window (Figure 15) to ensure data are in the location expected and that VBI and authoritative data sets are collocated.

Keeping the interface responsive while data discovery is processing is an important aspect of the program. This is accomplished using background concurrent threads to run the data discovery program and Qt slots and signals to pass information between the threads. Doing this avoids the appearance of crashing such as the window locking up and allows the user to monitor program progress.

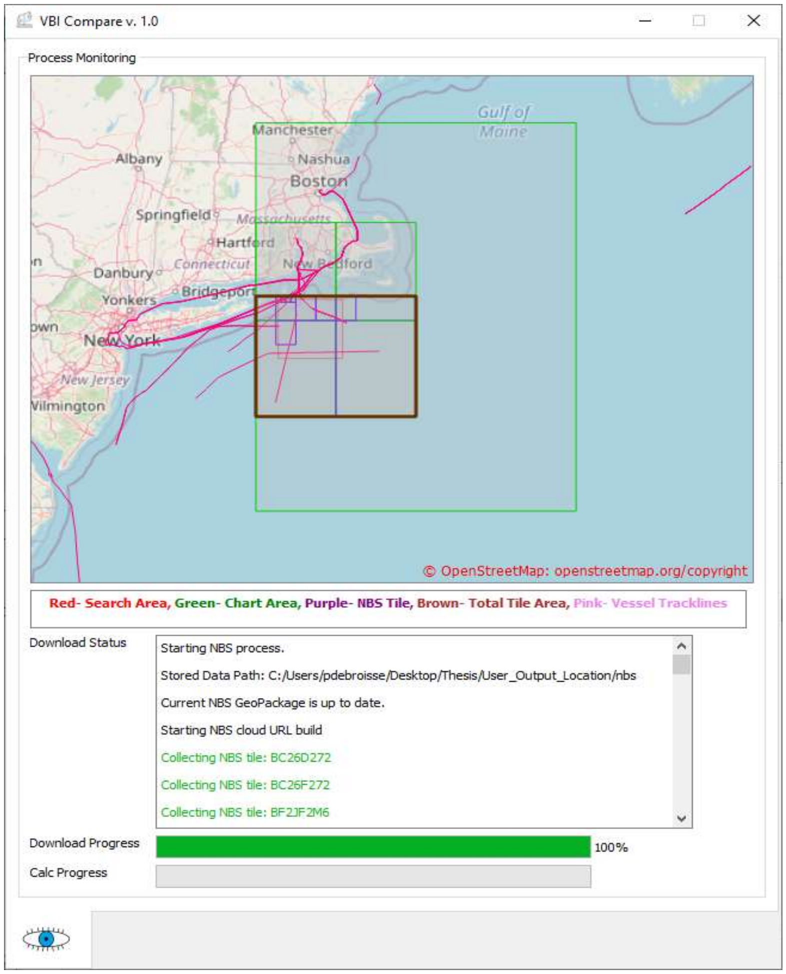


Figure 15: VBI Compare process monitoring window.

Within the VBI Compare monitoring window, the QGIS display is used to show chart outlines, NBS tile outlines, and vessel track lines. Additionally, if an area search is used, the geographic search area will also be displayed. As data discovery is happening in a background thread, the monitoring window displays any files that are discovered in the S3 buckets. It only displayed data that were actually available in the S3 buckets, not necessarily all the files suggested by the various API searches as there are occasions where files existing in the API server may not yet have an equivalent file created in the S3 bucket. This window zoomed out automatically upon completion to show all the data collected and could be panned and zoomed by the user manually as well.

As the GIS display is populated with data, a status update box is simultaneously updated with text status read outs. The user can read these statements to see how the program is progressing and what data has been collected and where it is located. A list of affected MCD ENC charts was also provided to the user in this window.

4 A Case Study

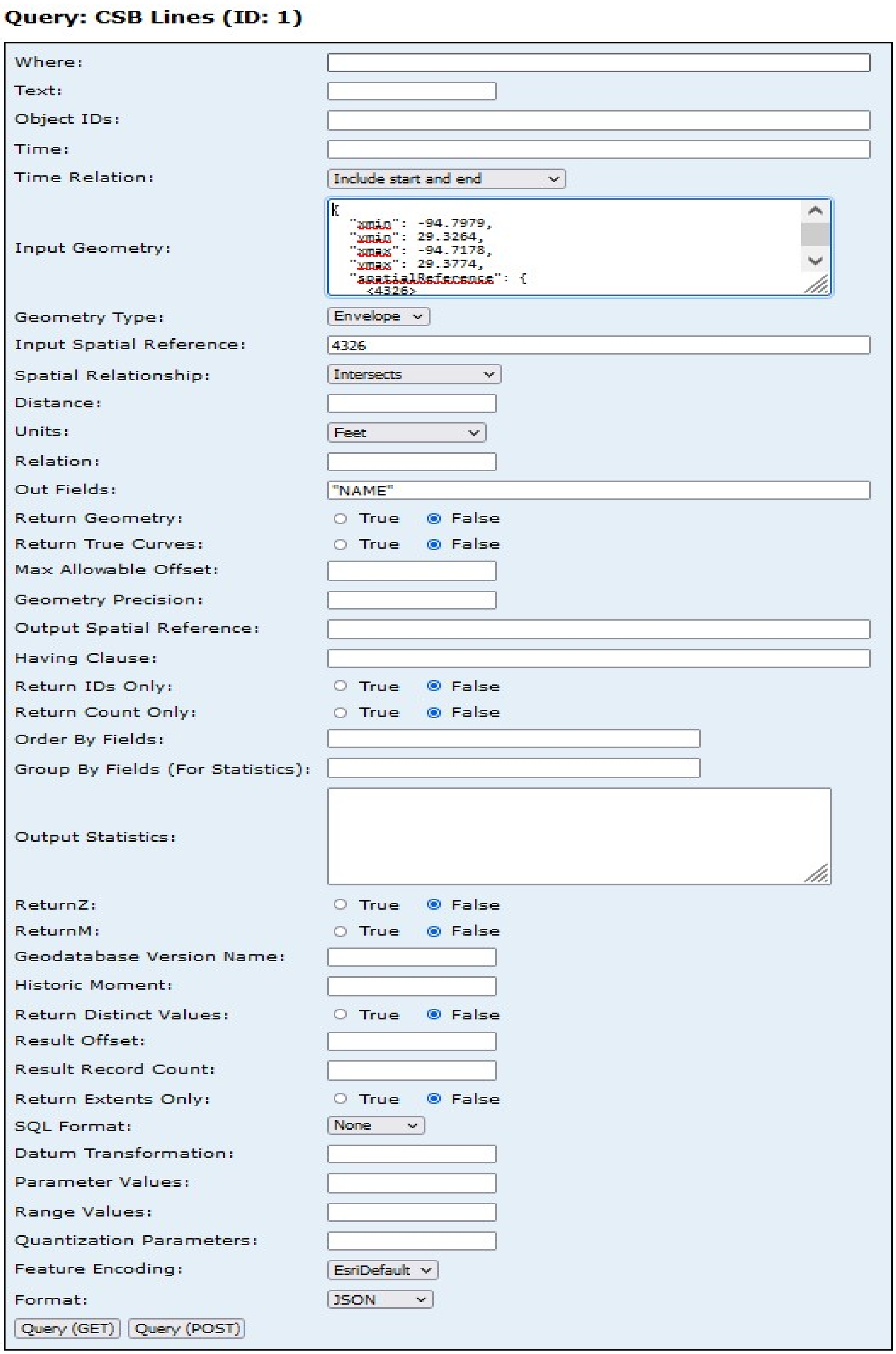
A case study using the Galveston Ship Channel to assess the application of VBI Compare as a data discovery and management tool was completed. For comparison, the same data discovery was attempted both manually and using VBI Compare. At the time of writing, NCEI is in the process of updating the IHO DCDB S3 database. For the last two years, DCDB has been restructuring its data pipeline to include pushing newly received data to their S3 bucket. This new operational pipeline has been built and is currently in the NOAA IT review and deployment process. In the meantime, a snapshot of data from March 2022 is available in the bucket for testing purposes. This update made a fully operational DCDB S3 bucket unavailable during the testing and demonstration phase. To continue testing, a method of bypassing the DCDB bucket search using previously downloaded CSV files was incorporated into the code until such time as the DCDB bucket becomes available again. One of those track lines is located in the Galveston Ship Channel which also has NBS data compiled. To complete a proof of concept, the search criteria used in VBI Compare were constrained to coincide with this area. VBI Compare was created as outlined in previous sections and information within the code has been updated with the future DCDB S3 bucket name and file conventions that will be used by DCDB such that the bypasses can be removed as soon as the DCDB updates have been released.

For this case study, use case 16 (Table 3) was used because it is a work flow that could be used often by an HO and exercises VBI Compare to its maximum capability. Use case sixteen searches the DCDB CSB S3 bucket first based on a user defined geographic area, and subsequently searches the NBS bucket using the resulting DCDB track line geometry. VBI Compare was instructed to collect S3 URLs for cloud-based processing. The manual data discovery and management attempt followed this same workflow.

4.1 Case Study Using Manual Methods

The first step of the manual workflow was to use the DCDB CSB ArcGIS REST API to determine the CSB files that are available in the area of interest. The settings used for the query can be seen in Figure 16. The query resulted in a JSON format text output listing one thousand file names. As stated before, this output does not provide data or links to data. To overcome this, the user copied the JSON text to a Microsoft Excel file. The JSON text was then parsed to yield only the file names. Using Excel string manipulation formulae, the user translated the file names into URLs for the data files in the DCDB CSB S3 bucket. Excel was used instead of, for example, Python, which may have been more efficient because Excel is a common data manipulation software and many users have had at least entry level training in its use. The DCDB portion of the workflow was complete when the user copied these URLs to a plain text file.

VBI Compare would collect the track line geometry from each of the lines found during this process. It would then use these to determine the NBS tiles that intersect the track lines. During testing, it was found that collecting and managing this track line geometry was difficult without writing a script. It became too cumbersome to attempt, so the user opted to query the NBS GeoPackage using the original user supplied area rather than track line geometry.



Query: CSB Lines (ID: 1)

Where:	
Text:	
Object IDs:	
Time:	
Time Relation:	Include start and end

Input Geometry:	{
	"xmin": -94.7979,
	"ymin": 29.3264,
	"xmax": -94.7178,
	"ymax": 29.3774,
	"spatialReference": {
	<4326>

Geometry Type:	Envelope
Input Spatial Reference:	4326
Spatial Relationship:	Intersects
Distance:	
Units:	Feet
Relation:	
Out Fields:	"NAME"
Return Geometry:	○ True ◉ False
Return True Curves:	○ True ◉ False
Max Allowable Offset:	
Geometry Precision:	
Output Spatial Reference:	
Having Clause:	
Return IDs Only:	○ True ◉ False
Return Count Only:	○ True ◉ False
Order By Fields:	
Group By Fields (For Statistics):	
Output Statistics:	
ReturnZ:	○ True ◉ False
ReturnM:	○ True ◉ False
Geodatabase Version Name:	
Historic Moment:	
Return Distinct Values:	○ True ◉ False
Result Offset:	
Result Record Count:	
Return Extents Only:	○ True ◉ False
SQL Format:	None
Datum Transformation:	
Parameter Values:	
Range Values:	
Quantization Parameters:	
Feature Encoding:	EsriDefault
Format:	JSON

Query (GET) Query (POST)

Figure 16: CSB ArcGIS API Query Settings

The user opened the NBS GeoPackage in the desktop QGIS application. After importing the GeoPackage, a layer with a shape file was created encompassing the user supplied search area. An intersection analysis was completed between the two files yielding a raster attribute table (RAT) listing the file names and links to two NBS tiles in the NBS S3 bucket (Figure 17). The user then copied the GeoTIFF links in the RAT to a plain text file, completing data discovery. The total time to complete manual data discovery took about twenty-one minutes, thirteen for the DCDB steps and eight for the NBS steps. If the user wanted to translate all the URLs collected from HTTP to S3 format like VBI Compare does, an additional eight minutes were required.

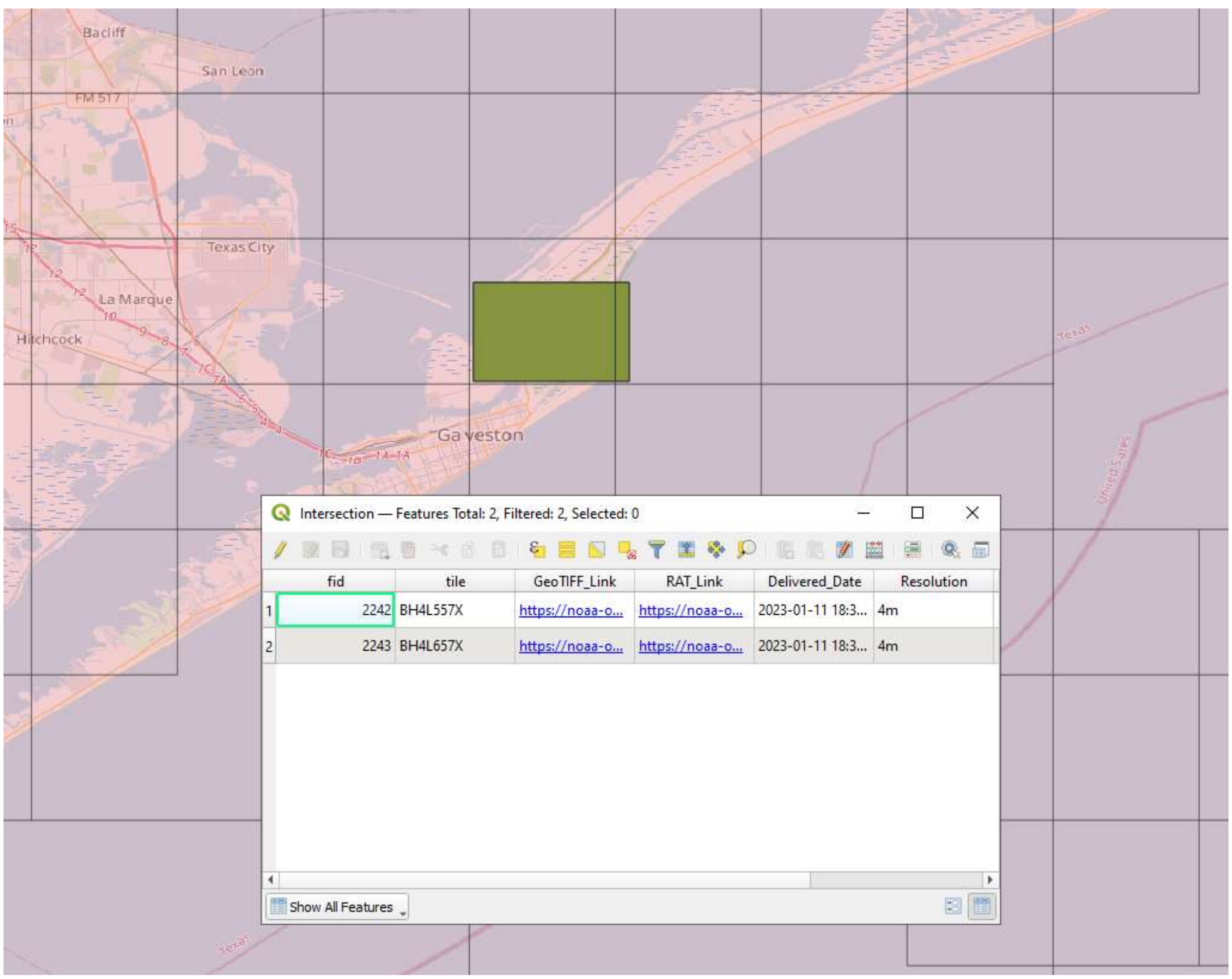


Figure 17: Manual NBS Query Results in QGIS

Several issues became apparent during this process shedding light on the limitations of manual data discovery. The first was that the DCDB API query does not result in any actual data or links to data. The need to translate the results from the query into URLs was a tedious and time-consuming process. While significant time might be saved with an experienced Excel user, the process is still tedious and error prone. Secondly, the user found it difficult to manage the track line geometry for each line and chose not to use it to inform the NBS query. While the results are still valid using this method, it collects far fewer NBS tiles. While it does ensure that the area of interest is covered, collecting all the tiles a track line crosses would result in more data points for the reputation algorithm leading to more accurate reputation scores. The QGIS intersection analysis did not take much time, but still required knowledge of the QGIS platform. The file management of the two plain text files was not particularly taxing, but in the case of local downloads it could be. The twenty-one minutes to complete this data discovery did not include the time to click each link and save each file in a manually constructed directory tree that would be required for local processing. While not fully evaluated, it was clear that the time added for local downloads and file management would be significant.

The user also skipped the step to determine the MCD charts that could be affected by a reputation calculation with the discovered datasets. To do so would require the collection of the track line geometry and an API search against that geometry using the MCD Rescheme Status API. There are six ENC resolution bands. Band one is excluded because of the small scales it uses, so each NBS tile would have to be cross referenced to each of the five remaining bands to determine the charts of interest. For one thousand track lines, this would result in five thousand manual API queries.

4.2 Case Study Using VBI Compare

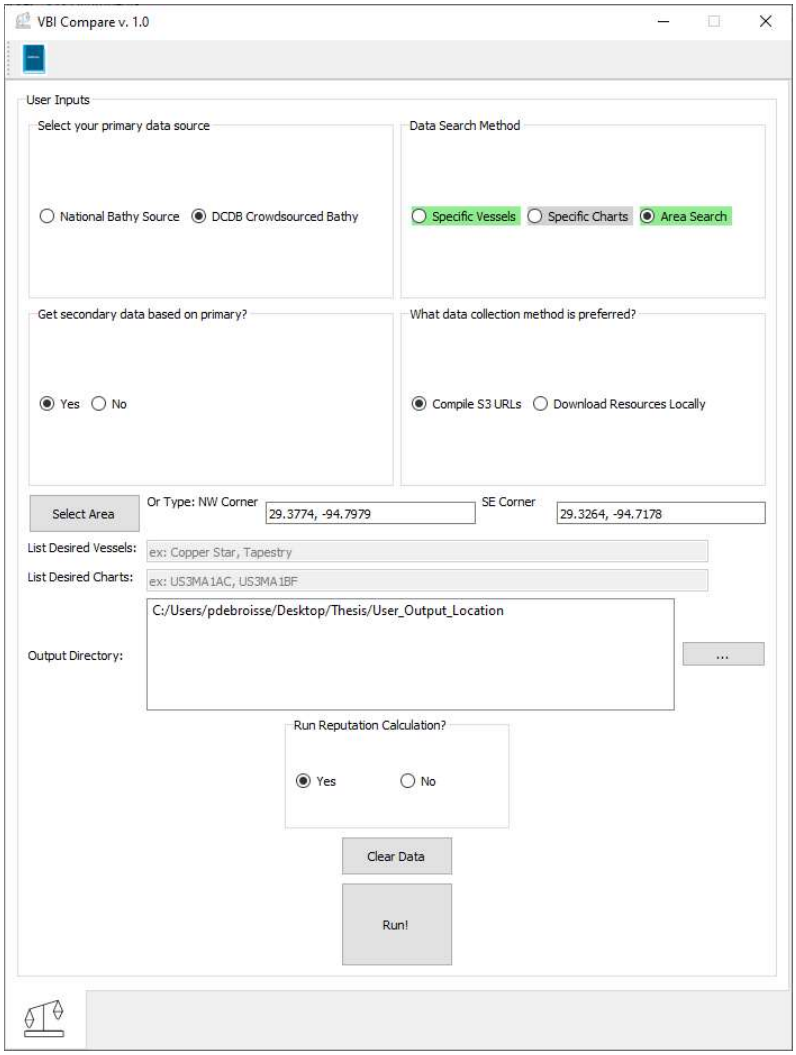


Figure 18: Case Study VBI Compare Parameters

The parameters used for this case study in VBI Compare were as shown in Figure 18. As a previously downloaded CSB file had to be used, a pre-constructed plain text file for the DCDB data was also used. The text file contained the local address to this downloaded CSV file. Under normal operation, both plain text files for the two types of data would have contained the same type of data, either S3 URLs or local addresses and would have been generated programmatically. The Calder VBI reputation and data uncertainty calculation algorithm can manage both data types simultaneously, thus this bypass is a viable option. The plain text file created by VBI Compare for the NBS data is the output that would normally be expected given the user inputs.

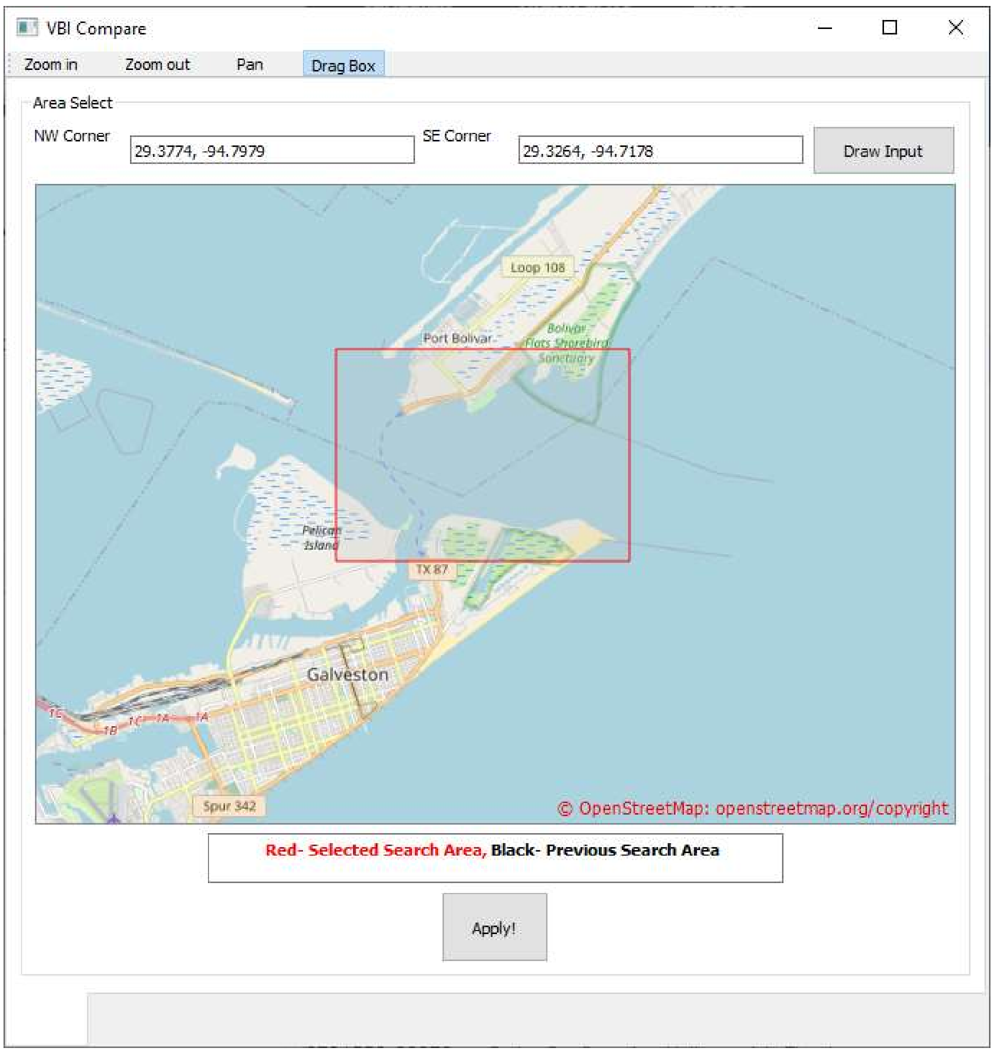


Figure 19: Case Study Area Selection

The area of interest was selected using the Select Area tool (Figure 19) by typing in the coordinates for the same bounding box as used in the manual test. The resulting area box was used for the entire case study.

4.2.1 VBI Compare Data Discovery

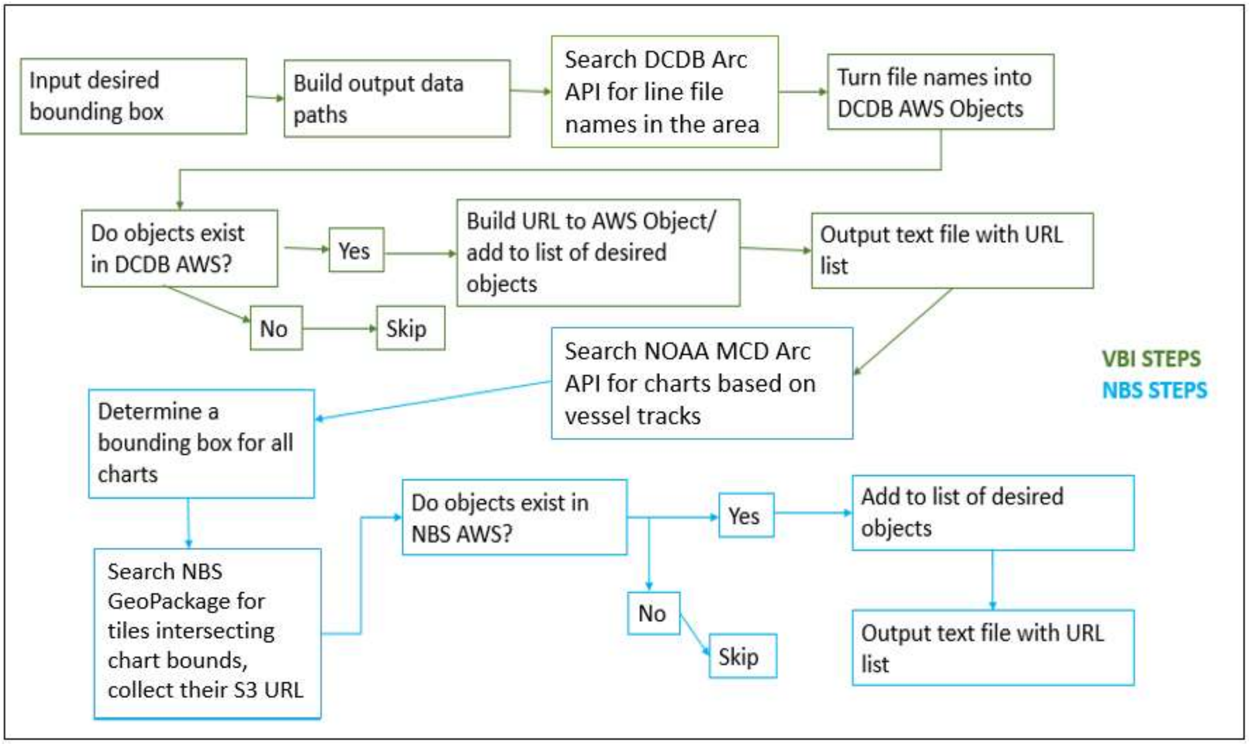


Figure 20: Case Study Programmatic Process

A visualization of the process followed for the use case of this case study is shown in Figure 20 above. VBI Compare started by creating the output directory structure for the various files the program would create. Using the supplied search area, it programmatically queried the DCDB API for any vessel track lines that intersect that area and collected the file names and geometry of those track lines. The file names were then automatically translated into S3 object names to search for in the DCDB S3 bucket. VBI Compare then searched the S3 bucket for the object names and collected the URLs to files that exist in the bucket. Once the DCDB S3 bucket search was complete, a text file containing all matching S3 URLs was saved.

VBI Compare then started the NBS search by using the track line geometries it collected to search the MCD ENC API to find all chart tiles that intersect them. Using the bounding boxes of the MCD charts it found, VBI Compare queried the NBS GeoPackage using SpatiaLite [40] to determine all NBS tiles that intersect them. The S3 bucket was then searched for these objects and, if they existed, the S3 URL to that object was collected from the GeoPackage file and set aside. Once the NBS S3 bucket search was complete, all the S3 URLs for NBS data were written into a plain text file and saved in the folder structure previously created.

The VBI Compare run of the case study resulted in the same one thousand DCDB track lines as were collected manually but completed this step in about thirty seconds. Additionally, the track line geometry was able to be collected and managed by the program unlike the manual test. Since track line geometry was compiled, 141 S3 URLs for NBS tiles were collected covering the entire length of each line. This step was completed in approximately one and a half minutes. The program resulted in the creation of two plain text files in the constructed directory tree as well as a download of the most up to date NBS GeoPackage file. The user also had an older GeoPackage file in the directory which VBI Compare automatically moved to a constructed archive folder. Finally, the program took thirteen seconds to determine that 122 MCD charts could be affected by the reputation calculations using the data sets collected. It also provided the user with a list of those chart numbers. In total, it took VBI Compare three minutes to collect, package, store, and display all this data (Figure 21). This test was repeated several times with varying loads on the internet bandwidth. These tests all fell within a total runtime range of two minutes and thirty seconds to three minutes and thirty seconds. This range of results showed some dependence on internet bandwidth, but was still much faster than the manual workflow, regardless of the load.

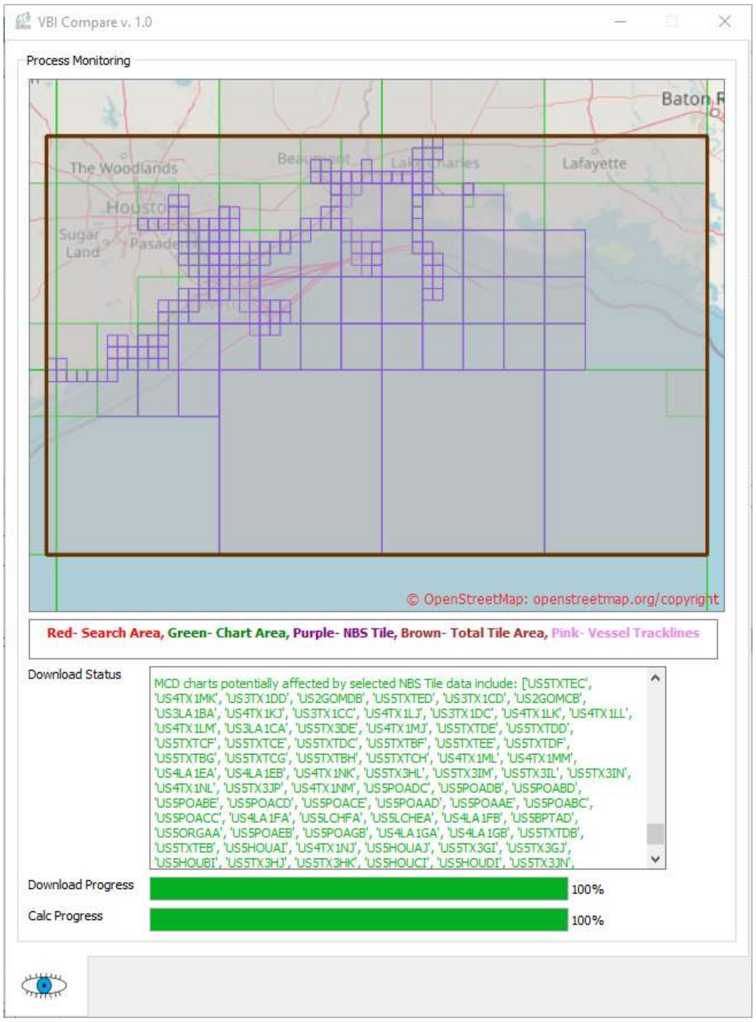


Figure 21: Case Study VBI Compare Results

4.3 Reputation Calculation

After data discovery is complete, the plain text files need to be supplied to the Calder reputation and data uncertainty calculation algorithm. VBI Compare does this using a Windows batch (.bat) file. An output folder structure is created to store the batch file and the output files from the calculation algorithm. The NBS and VBI plain text files that contain the S3 URLs are provided to the calculation algorithm via this batch file.

The file also contains step by step instructions for the algorithm to follow as well as any inputs or outputs these steps require.

The outputs from the calculation algorithm are saved to the folder structure VBI Compare created alongside the batch file. Additionally, VBI Compare generates a figure and stores it in the same folder as the other reputation output files. This figure shows the reputation and uncertainty of the vessel as it changes through the vessel's transit. Figure 22 shows the output figure from this case study.

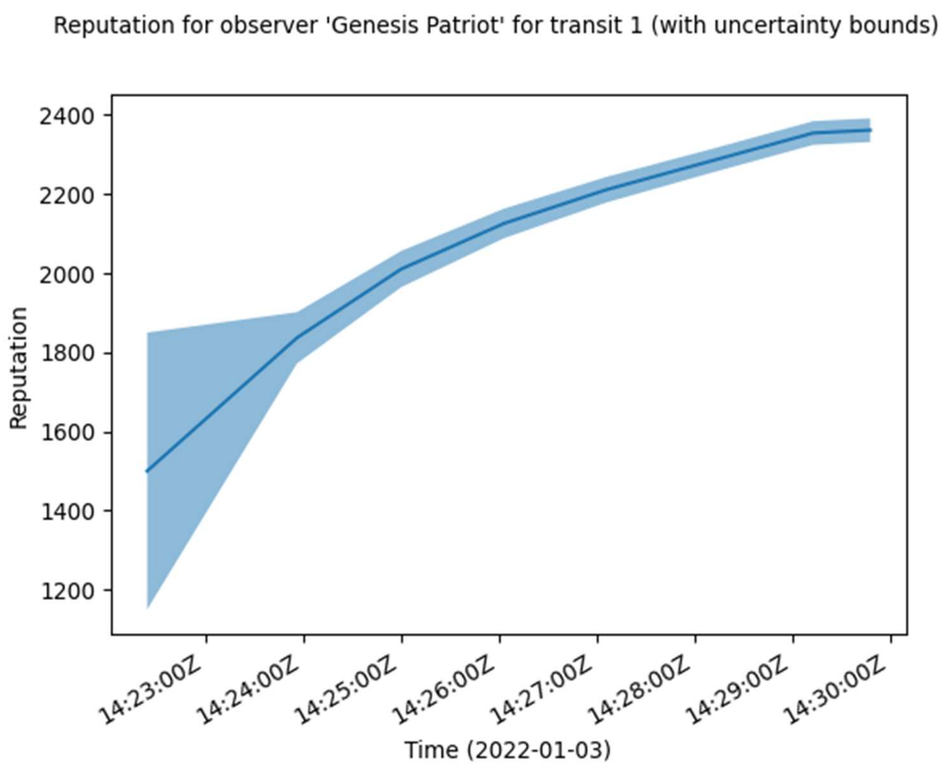


Figure 22: Example output from VBI reputation calculation.

5 Discussion, Conclusions, and Future Development

5.1 Discussion

The uptake of VBI as a data source by hydrographic offices has met resistance due to the time expense of determining data quality or constructing a final usable product. It has been shown here that VBI Compare in conjunction with a data processing algorithm could unlock VBI as a viable data source by significantly reducing the workload of determining data quality. While VBI Compare has significant potential, additional updates to the DCDB and the NBS could further these contributions.

As shown, the DCDB S3 bucket is not directly searchable necessitating the use of its API and a translation of file names from API to S3 bucket convention. This could be alleviated by either making the file naming conventions of both databases the same, or alternatively using a GeoPackage type file within the S3 bucket similar to that of the NBS. Further, certain metadata are available in the GeoJSON files accessed via the API that are not available in the CSV files of the S3 bucket. Specifically, vessel characteristics would be helpful for users processing this data. If the DCDB were to make a dictionary file in the S3 bucket containing the vessel characteristics associated to a specific Unique User ID (UUID) or include them in the CSV files for each line, VBI Compare would be able to collect this data as part of its data discovery process.

The naming convention of the data tiles in the NBS is different from those of the ENC chart tiles, making a translation necessary to determine which charts may be affected by a data quality calculation. It also makes searching by a specific ENC chart more difficult. The GeoPackage file used by NBS could also serve as a cross reference file for nautical charts. If a data column existed in the file which held a list of coincident

ENC charts for each NBS tile, then an intersection analysis could be executed on the GeoPackage by both geometry and chart number removing the need to utilize the MCD ENC Rescheme API.

5.2 Conclusions

VBI Compare has made several contributions towards the expansion of the use of VBI by HOs for authoritative or monitoring purposes. This case study showed that rapid discovery of collocated authoritative and VBI data sets is possible using a semi-automated programmatic method. Not only can data sets be collected rapidly, but they can also be discovered in any geographic area where authoritative and VBI data sets exist. Additionally, by using the open-source Python language, VBI Compare can be adapted to the needs of the user allowing for different data sources or search criteria in the future. Furthermore, its design as a desktop application allows it to be integrated into existing hydrographic tool suites for ease of dissemination.

It was shown in this case study that consistent results given the same input were achieved, ensuring that VBI Compare could be used in further testing and developments of the VBI reputation and data uncertainty calculation algorithms. Its robust operation, simple interface, and transportability make VBI Compare, in conjunction with the Calder reputation and uncertainty calculation algorithm, a viable alternative to calculating VBI reputation and uncertainty when compared to current manual processing.

5.3 Future Development

Due to the updates being built into DCDB and the resulting inaccessibility, some development and testing of VBI Compare remains. Most importantly, the file naming convention and S3 bucket name have been updated by DCDB. These changes have been preemptively coded into VBI Compare, but it needs to be assessed to ensure all file name capture and data discovery is still functional once DCDB makes its S3 bucket available again[i]. Additionally, there is a desire to allow VBI Compare to collect the data for each vessel provided by the user individually, rather than all at once. VBI Compare does this when provided a list of charts but the inaccessibility to DCDB limited the ability to develop and assess this feature for vessels. Adding this capability will make it easier to associate specific calculation files to specific vessels. Once the DCDB S3 bucket is available, developers can add the ability to complete the comparison for each vessel provided in a list using the existing functionality for a given list of charts as a reference.

The National Bathymetric Source was the selected authoritative database for this case study, but it may not be the desired option for all users. Further development by end users or by the VBI Compare team could add options for other authoritative data sources such as the Canadian NONNA [41] data set. Similarly, VBI Compare is built specifically to collect VBI data from the DCDB CSB server, but users may have a desire to compare other data sources to an authoritative source. Options might include external source multibeam data or authoritative data collection sources. The Calder reputation and uncertainty calculation algorithm can produce comparisons for these other data sources, so VBI Compare could be expanded to perform data discovery for these data types as well. Additional GUI tabs could be added to VBI Compare to manage these types of requests. Programmers could then mimic the CSB data capture portion of VBI Compare for their data set of interest.

Currently, VBI Compare stores VBI data by vessel name. It is possible that vessels may have the same name, have different spellings of their name in the database, or change their name causing confusion. An alternative could be to use vessel UUIDs. In the DCDB, each vessel should have its own UUID regardless of spelling or name change. The API can be queried by UUID, and they are stored within each CSV file in the S3 bucket. Using these UUIDs could make VBI Compare more robust by accounting for these vessel naming challenges. UUIDs are less human readable than the vessel names themselves and not easily determined, so a dictionary file cross referencing UUID to platform name would make this an easier prospect. The use of UUIDs would also allow for vessels tagged “anonymous” to be separated from each other.

The current version of VBI Compare is designed specifically to use the current S3 buckets and naming conventions of the NBS and DCDB. If these were to change, VBI Compare’s functionality would erode. VBI Compare is modular by using defined functions to execute the various sub-steps of the program. Additional functions could be written that would allow the program to interact with databases not on the AWS S3 platform. The functions could be in addition to, or replacements for current functions. Further, the ability to take in a configuration file could be added so that the user could supply the database locations and naming conventions to the program with each run allowing for more flexibility once functions are added for use with other database types.

Additionally, because the program is open source, each individual user has the option to adjust or rewrite any of the functions to fit their specific needs.

Refinement of the VBI Compare code could result in increased efficiency. Further, additional query constraints such as a date range, or the ability to combine vessel or chart name with geographic area, could be integrated to give the user further options.

---

[i] On 27FEB 2023, the DCDB released an update to their AWS S3 bucket and data handling procedures. Since the writing of this thesis, updates have been made to the Python code for VBI Compare to integrate these updates. This has made the need for the bypass used in the case study described above obsolete.

References

[1] National Centers for Environmental Information, "Uses of hydrographic Survey Data," National Oceanic and Atmospheric Administration, [Online]. Available: https://www.ngdc.noaa.gov/mgg/bathymetry/noshdb/nosuses.HTML. [Accessed 24 02 2022].

[2] D. J. Trump, *Memorandum on Ocean Mapping of the United States Exclusive Economic Zone and the Shoreline and Nearshore of Alaska,* Washington D.C.: United States of America, 2019.

[3] Integrated Ocean & Coastal Mapping, "Status of Seafloor Mapping Within U.S. Waters," National Oceanic and Atmospheric Administration, March 2023. [Online]. Available: https://iocm.noaa.gov/seabed-2030-status.html. [Accessed 15 March 2023].

[4] Great Lakes Observing System, "Lakebed 2030," Great Lakes Observing System, 2021. [Online]. Available: https://glos.org/priorities/lakebed-2030/. [Accessed 06 Oct 2022].

[5] Seabed 2030 Project, "Mapping Progress," British Oceanographic Data Centre, 2022. [Online]. Available: https://seabed2030.org/mapping-progress. [Accessed 16 12 2022].

[6] United Nations, "Blue Economy Definitions," [Online]. Available: https://www.un.org/regularprocess/sites/www.un.org.regularprocess/files/rok_part_2.pdf. [Accessed 25 March 2022].

[7] National Oceanic and Atmospheric Administration, "NOAA Blue Economy Strategic Plan 2021-2025," 19 January 2021. [Online]. Available: https://aambpublicoceanservice.blob.core.windows.net/oceanserviceprod/economy/Blue-Economy%20Strategic-Plan.pdf. [Accessed 25 March 2022].

[8] Seabed 2030, "About the Seabed 2030 Project," British Oceanographic Data Centre, 2022. [Online]. Available: https://seabed2030.org/about-us. [Accessed 25 March 2022].

[9] K. Mizokami, "15 Years Ago, a U.S Navy Submarine Ran Into a Mountain," 09 January 2020. [Online]. Available: https://www.popularmechanics.com/military/navy-ships/a24158/uss-san-francisco-mountain-incident/. [Accessed 04 April 2022].

[10] S. LaGrone, "Investigation Concludes USS Connecticut Grounded on Uncharted Seamount in South China Sea," USNI News, 01 November 2021. [Online]. Available: https://news.usni.org/2021/11/01/investigation-concludes-uss-connecticut-grounded-on-uncharted-sea-mount-in-south-china-sea. [Accessed 23 March 2022].

[11] L. Carter, D. Burnett, S. Drew, G. Marle, L. Hagadorn, D. Bartlett-McNeil and N. Irvine, "Submarine Cables and the Oceans - Connecting the World," *UNEP-WCMC BIodiversity Series No.31,* 2009.

[12] S. Helmy and A. Shrabia, "Mega Container Ships, Pros, Cons and Its Implication," *Journal of Shipping and Ocean Engineeri,* vol. 6, pp. 284-290, 2016.

[13] National Ocean Service, "An Inch of Water. What's it Worth?," National Oceanic and Atmospheric Administration, [Online]. Available: https://oceanservice.noaa.gov/economy/inch-water/. [Accessed 11 04 2022].

[14] National Ocean Service, "Precision Marine Navigation," National Oceanic and Atmospheric Administration, 02 Mar 2022. [Online]. Available: https://oceanservice.noaa.gov/navigation/precision-navigation/. [Accessed 16 Dec 2022].

[15] R. Ward, L. Alexander and B. Greenslade, "IHO S-100: The New IHO Hydrographic Geospatial Standard for Marine Data and Information," *International Hydrographic Review,* vol. No. 1, pp. 44- 55, 2009.

[16] International Hydrographic Organization, "S-100 Universal Hydrographic Data Model," International Hydrographic Organization, 24 May 2022. [Online]. Available: https://iho.int/en/s-100-universal-hydrographic-data-model. [Accessed 16 Dec 2022].

[17] R. J. Weaver and D. N. Slinn, "Influence of Bathymetric Fluctuations on Coastal Storm Surge," *Coastal Engineering,* vol. 57, no. 1, pp. 62-70, 2010.

[18] J. Jencks and A. Chappell, "Working Together to Map the Gaps," *Hydro International,* pp. https://www.hydro-international.com/content/article/working-together-to-map-the-gaps, 11 SEP 2018.

[19] International Maritime Organization, "International Conventino for the Safety of Life at Sea (SOLAS), 1974," 2019. [Online]. Available: https://www.imo.org/en/About/Conventions/Pages/International-Convention-for-the-Safety-of-Life-at-Sea-(SOLAS),-1974.aspx. [Accessed 16 Dec 2022].

[20] International Hydrographic Organization, "Guidance on Crowdsourced Bathymetry," International Hydrographic Organization, Monaco, 2020.

[21] International Hydrographic Organization, "S-44 Edition 6.0.0," Sep 2020. [Online]. Available: https://iho.int/uploads/user/pubs/standards/s-44/S-44_Edition_6.0.0_EN.pdf. [Accessed 16 Dec 2022].

[22] S. Wiehle, B. Martinez, K. Hartmann, M. Verlaan, T. Thronton, S. Lewis and D. Schaap, "The BASE-platform project: Deriving the bathymetry from combined satellite data," in *Proceeding of Hydro 2016*, Warnemunde, 09NOV2016.

[23] A. R. Klemm and T. Krabiel, "From Crowd to Chart: Methods and Applications of Crowdsourced Bathymetry in Support of Safe Navigation," MAR 2023. [Online]. Available: https://github.com/anthonyklemm/Crowdsourced_Bathy_Processing/blob/main/Klemm_CSB_USHYDRO23.pdf. [Accessed 06 APR 2023].

[24] K. Jerram and V. Schmidt, "Object Detection with Phase-Measuring Bathymetric Sidescan Sonar Depth Data," Center for Coastal and Ocean Mapping, Durham, NH.

[25] B. Calder, "Estimating Observer and Data Reputation in Mariner Volunteered Bathymetry," *International Hydrographic Review,* vol. 26, 2021.

[26] Office of Coast Survey, "BLUETOPO A curated collection of high resolution seafloor models," National Oceanic and Atmospheric Administration, [Online]. Available: https://www.nauticalcharts.noaa.gov/data/bluetopo.html. [Accessed 16 Dec 2022].

[27] Amazon, "Amazon S3," Amazon Web Service, 2023. [Online]. Available: https://aws.amazon.com/s3/. [Accessed 29 MAR 2023].

[28] HydrOffice team, "HydrOffice," HydrOffice team, 2022. [Online]. Available: https://www.hydroffice.org/. [Accessed 16 Dec 2022].

[29] Office of Coast Survey, "Introduction to Pydro," National Oceanic and Atmospheric Administration, [Online]. Available: https://svn.pydro.noaa.gov/Docs/html/Pydro/universe_overview.html. [Accessed 16 Dec 2022].

[30] GeoJSON Working Group, "GeoJSON (RFC 7946)," Internet Engineering Task Force, 2016.

[31] IHO Data Centre for Digital Bathymetry, "csb (MapServer)," National Center for Environmental Information, [Online]. Available: https://gis.ngdc.noaa.gov/arcgis/rest/services/csb/MapServer. [Accessed 19 Dec 2022].

[32] Y. Safranovich, "Common Format and MIME for CSV Files," The Internet Society, 2005.

[33] Open Geospacial Consortium, "OGC GeoTIFF standard," 14 Sep 2019. [Online]. Available: http://docs.opengeospatial.org/is/19-008r4/19-008r4.html. [Accessed 16 Dec 2022].

[34] National Geodetic Survey, "North American Vertical Datum of 1988 (NAVD 88)," National Oceanic and Atmospheric Adminisration, 12 Jul 2018. [Online]. Available: https://geodesy.noaa.gov/datums/vertical/north-american-vertical-datum-1988.shtml. [Accessed 13 Jan 2023].

[35] Center for Operational Oceanographic Products and Services, "Tidal Datums," National Oceanic and Atmospheric Administration, [Online]. Available: https://tidesandcurrents.noaa.gov/datum_options.html. [Accessed 13 Jan 2023].

[36] Office of Coast Survey, "BlueTopo Data Specifications," National Oceanic and Atmospheric Administration, [Online]. Available: https://www.nauticalcharts.noaa.gov/data/bluetopo_specs.html. [Accessed 20 9 2022].

[37] Open Geospatial Consortium, "OGC GeoPackage Encoding Standard," Open Geospatial Consortium, 2021. [Online]. Available: https://www.geopackage.org/spec130/. [Accessed 16 Dec 2022].

[38] Office of Coast Survey, "MarineChart_Services/Status_New_NOAA_ENCs (MapServer)," National Oceanic of Atmospheric Administration, [Online]. Available: https://gis.charttools.noaa.gov/arcgis/rest/services/MarineChart_Services/Status_New_NOAA_ENCs/MapServer. [Accessed 19 Dec 2022].

[39] Office of Geomatics, National Geospatial-Intelligence Agency, 04 Oct 2022. [Online]. Available: https://earth-info.nga.mil/index.php?dir=wgs84&action=wgs84. [Accessed 16 Dec 2022].

[40] A. Furieri, "SpatiaLite," Regione Toscana - Settore Sistema Informativo Territoriale ed Ambientale, 01 NOV 2020. [Online]. Available: https://www.gaia-gis.it/fossil/libspatialite/index. [Accessed 15 MAR 2023].

[41] Canadian Hydrographic Service, "Canadian Hydrographic Service Non-Navigational (NONNA) Bathymetric Data," Government of Canada, 21 APR 2022. [Online]. Available: https://open.canada.ca/data/en/dataset/d3881c4c-650d-4070-bf9b-1e00aabf0a1d. [Accessed 05 APR 2023].

[42] Seabed 2030 Project, "Mapping Progress," The Nippon Foundation-GEBCO, 2022. [Online]. Available: https://seabed2030.org/mapping-progress. [Accessed 23 Mar 2022].

[43] The QT Company, "QT for Python," The QT Company, 2022. [Online]. Available: https://www.qt.io/qt-for-python. [Accessed 27 SEP 2022].

[44] QGIS Project, "Documentation for QGIS 3.22," QGIS Project, 2022. [Online]. Available: https://docs.qgis.org/3.22/en/docs/index.html. [Accessed 27 SEP 2022].

[45] J. Bartolotta, "NOAA Teacher at Sea Blog," National Oceanic and Atmospheric Administration, 05 JUN 2019. [Online]. Available: https://noaateacheratsea.blog/tag/multibeam-sonar/. [Accessed 04 OCT 2022].

[46] L. Gupta, "HTTP Methods," Rest API Tutorial, 11 DEC 2021. [Online]. Available: https://restfulapi.net/http-methods/. [Accessed 28 NOB 2022].

[47] International Hydrographic Organization, "Data Centre for Digital Bathymetry Viewer," National Center for Environmental Information, 20 Oct 2022. [Online]. Available: https://www.ncei.noaa.gov/maps/iho_dcdb/. [Accessed 16 Dec 2022].

[48] R. T. Fielding, M. Nottingham and J. Reschke, "HTTP Semantics," Internet Eningeering Task Force (IETF), June 2022. [Online]. Available: https://httpwg.org/specs/rfc9110.html#GET. [Accessed 19 Dec 2022].