iDEM
integrator of Digital Elevation Models

by

Peter Bertil Johan Salomonsson
B.Sc., Victoria University of Technology, 1995

A Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

in the Department of Computer Science

iDEM
integrator of Digital Elevation Models

by

Peter Bertil Johan Salomonsson
B.Sc., Victoria University of Technology, 1995

Supervisory Committee

Dr. Yvonne Coady, Supervisor
(Department of Computer Science)

Dr. Sue Whitesides, Departmental Member
(Department of Computer Science)

Dr. Tania Lado Insua, Additional Member
(Ocean Networks Canada)

Supervisory Committee

Dr. Yvonne Coady, Supervisor
(Department of Computer Science)

Dr. Sue Whitesides, Departmental Member
(Department of Computer Science)

Dr. Tania Lado Insua, Additional Member
(Ocean Networks Canada)

ABSTRACT

Digital Elevation Models (DEM) are typically created through a variety of multi-step processes that are generally labour intensive. This thesis explores the trade-offs involved in automating these processes in order to produce a DEM at various resolutions, while minimizing artifacts and highlighting areas where artifacts or uncertainty may have been introduced.

The *iDEM* system is a prototype design to automate the creation of customized DEM complete with a detailed audit-trail of metadata history. Originally conceived as a solution to creating DEM for tsunami modelling, iDEM is applicable to modelling any spherical surface. The proposed framework is highly generalizable in that it leverages existing applications in a plug-and-play manner, essentially integrating them into a new system. The creation of DEM in our prototype design utilizes an amalgamation of three existing fusion methods that allow tessellation without edge distortion and propagates data uncertainty for every DEM generated. The challenge of integrating data in different formats is tackled by automatically generating customized DEM based on the selection of any module providing data fusion techniques applied to the best measurements available.

Contents

List of Tables

List of Figures

ACKNOWLEDGEMENTS

This work has been assisted by the mentoring and support of:

Y. Coady, PhD Associate Professor, Computer Science, University of Victoria

S. Whiteside, PhD Professor, Computer Science, University of Victoria

S. W. Neville, PhD Associate Professor, Electrical and Computer Engineering, University of Victoria

T. L. Insua, PhD, Ocean Analytics Program Manager, Innovation Centre, Ocean Networks Canada

S. Salomonsson, CPA CA, SDB CA PC

R. Jenkyns, Data Stewardship Team Lead, Innovation Centre, Ocean Networks Canada

K. Douglas, GIS Specialist - Digital Infrastructure, Innovation Centre, Ocean Networks Canada

Chapter 1

Introduction

Maps in the form of digital elevation models (DEM) can cover both land (topography) and the bottom of bodies of water, such as the ocean floor (bathymetry). DEM can be created through a variety of different multi-step processes that are generally labour intensive, requiring careful manual tracking of techniques applied, and possible errors introduced. This thesis explores the design of a more automated approach to produce DEM at various resolutions by combining both topography and bathymetry data while minimizing artifacts and highlighting areas where artifacts may exist.

1.1 Digital Elevation Models with an Audit Trail for Uncertainty

DEM can be utilized for tasks as diverse as archaeology, flight simulation, 3D visualization, and modelling water flow for hydrology. A larger DEM is often created by stitching together smaller existing DEM. Such stitching often results in inaccuracies known as *artifacts*. In the case of hydrodynamic modelling, artifacts can alter the behaviour of the water flow making the model unrealistic. Many attempts have been made during the past few years to smooth over such artifacts with varying success. In most cases the resulting DEM may be smoother with less visible artifacts but often smoothing masks the artifacts.

There are several common factors that tend to result in join distortions when existing DEM are combined into a larger model:

- Lack of data distortion caused by the sparsity of elevation data from which the DEM were originally created. Lack of data can cause significant deviation from reality.

- Data quality distortion caused by the quality of the elevation data from which the DEM were originally created. Measurement errors can cause significant deviation from reality [58].

- Resolution distortion caused by the differing resolutions of existing DEM. The join where a detailed high resolution model abuts a smoother low resolution model is an area of distorted transition between the two models [56].

- Processing distortion, many of the more sophisticated algorithms that are used to create the DEM inherently produce an output with greater inaccuracy at the edges [112].

These factors often manifest themselves in the combined larger DEM as artifacts that represent non-existent features, especially along the DEM joins. The process of correcting for random and systematic errors in raw elevation data is known as Cleaning. Ideally the larger DEM should be re-gridded and re-interpolated from cleaned elevation data covering the area of interest. Such re-interpolation would eliminate artifacts caused by joining DEM and reduce the propagation of any errors that may be inherent in the DEM.

Sometimes DEM are available for areas where no raw data can be obtained. Often existing DEM have a lower level of accuracy than the raw datasets due to extensive preprocessing. In other instances additional elevation data may be available to enhance an existing DEM but the original elevation data may not be available. In order to re-interpolate for a new area or to enhance an existing DEM a system is required that processes existing DEM as though they are merely a collection of preprocessed elevation points. The system would merge the new elevation points with the existing DEM elevation points and interpolate a new DEM based on the larger merged dataset. A method should be chosen to merge the datasets while minimizing artifacts that can result from the merging process.

The uncertainty of the data points in the various datasets can be utilized as part of the fusion process to determine the composition of the blended dataset. That same uncertainty can be carried through the interpolation process to create not only a DEM, but an accompanying model to capture the inherent uncertainty involved. We refer to this as the Digital Certainty Model (DCM), used in highlighting the relative accuracy of each grid point of the DEM. Here the term *certainty* is used to qualify a relative degree of trust. A DCM is an array that mirrors a DEM but this array is populated by uncertainty values directly associated with elevation data in the DEM. A DCM can be used to highlight areas where the model of the terrain may be inaccurate. As such, a DCM assists a user of a DEM visually, or numerically, helping assess the uncertainty or the quality of the DEM. Such assessment can allow a user to develop a level of trust in the DEM.

This research has been driven by Ocean Networks Canada's (ONC) need for extensive, regular, rectangular DEM, at various resolutions, of the waters off the coast of British Columbia, Canada. The proposed system, and output DEM, need to be accessible to researchers without expert knowledge of DEM creation. The resulting DEM should be accompanied by an audit trail justifying strengths and highlighting weaknesses. The proposed *integrator of Digital Elevation Models* is generally applicable to the creation of regular rectangular DEM with uncertainty anywhere.

Similar DEM can be utilized for most DEM dependent tasks not only within ONC but for any other user. DEM dependent tasks include (but are not limited to) navigation, current flow, hydrology, flooding/storm surge, creation of relief maps, 3D visualization, effect of rising sea levels, underwater acoustics, terrain analysis, flight simulation, archaeology, etc. The DEM required by these tasks have customized water level references based on the particular application. The proposed design of iDEM in this thesis has been developed based on the tsunami modelling case. An accompanying analysis is presented of hardware disk space requirements.

The design of iDEM is based on the requirements that system should be:

1. partially or fully automated;
2. store all input data (cleaned elevation points, existing DEM, and raw elevation points);
3. create and store DEM at the desired location and at the desired resolution;
4. create and store a DCM of each DEM created;
5. allow the re-creation or enhancement of any existing DEM in the system;
6. track metadata of all input, processing and output; and
7. minimize development costs, time-lines, and system maintenance.

1.2 Proposed System

The design presented here for an integrator of Digital Elevation Models (iDEM) can best be summarized as:

1. Utilizing existing robust, free, Open Source applications and utilities, which:
   (a) minimize initial costs;
   (b) leverage external resources for future development and maintenance; and
   (c) reduce development time-lines.
2. Utilizing a relational database as the core component. This allows central storage of:
   (a) the various sets of input data;
   (b) metadata;
   (c) the resulting DEM;
   (d) the accompanying uncertainty models; and
   (e) cross-referencing indices enabling justification of the output.
3. Automating the data fusion process with a simple ad hoc data merging algorithm as a starting point for future replacement by a rigorous method:
   (a) ad hoc data merging is based on a sorting method to determine which data points to keep and which to discard;
   (b) the sorting method is based on inherent characteristics of the input data, primarily density and/or uncertainty;
   (c) the sorting method is applied utilizing two merging methods that are complimentary with respect of the relative resolutions of the dataset being merged; and
   (d) the appropriate complimentary merging method is automatically chosen by an adjustable resolution similarity algorithm.
4. Integrating and enhancing best practices as identified by a literature review in order to:
   (a) process high density data into a lower density grid with uncertainty;

   (b) optionally reduce sampling errors by strategically aggregating lower density data; and
   (c) process lower density data by interpolation with uncertainty.

5. Enfolding the relational database and integrated applications in python wrappers for interfacing with a python framework for:
   (a) automating the creation of the database structure;
   (b) partially automating the data input;
   (c) partially automating data preprocessing; and
   (d) integrating and automating existing applications for processing the data.

The proposed design for the iDEM system produces digital elevation models with minimal user interaction. The automation of the process allows users to concentrate on those tasks that benefit most from their attention.
There are *three* claims resulting from this work:

1. Significant portions of what are typically manual, labour-intensive processes can be automated for the creation of customized DEM.
2. The framework allows functions and algorithms implemented in existing applications to be re-purposed.
3. An audit trail and uncertainty model of the resulting DEM can be systematically developed to promote trust in the final product.

1.2.1 The Significance of the iDEM Framework

Data homogenisation, uncertainty estimation, data fusion, and digital certainty models for DEM creation are not new. Many applications exist that allow the user to interactively process elevation data. Some applications can provide uncertainty estimation. In most existing applications data fusion is largely a user intensive task. Some applications do allow for the creation of digital certainty models.

The existing applications are often feature rich, extremely flexible, user controlled and intensely user interactive. The iDEM system design will have minimal user interaction. It is not designed for flexibility but is tailored to automate the storage and production of digital elevation maps with a full history of each DEM creation process. A DCM will automatically accompany each DEM.

The iDEM framework reduces the need for expert training, some understanding is required to input data into the system. It raises the importance of the metadata by requiring dataset registration in order that the application is aware of data nuances.

Many of the individual subsystems proposed to be integrated into the iDEM framework already exist. Most of them are considered powerful applications in their own right. Each application contributes one or more significant functions as subsystems of this new system. It is the overall framework binding the integrated applications together that is the achievement of this work.

Some very important positive consequences arise from the the above claims. It is these consequences that comprise a significant positive contribution to research in the field of DEM creation.

Claim 1 *automatic creation of customized DEM* implies that DEM creation should not require expert training. Rather it should allow the DEM user to concentrate on those activities for which the DEM is required and ensure that each DEM created adheres to the same standards. The automated system should also result in the population of a central data repository that can allow the creation of DEM to leverage all existing datasets registered with iDEM.

Claim 2 *new framework to leverage existing applications* implies that the integrated functions should be able to leverage the development and rely on existing function verification performed on the integrated applications by the developers. Testing and verification of the framework should be able to be restricted to the integration system itself. It also implies that no matter which implementation of a function is utilized the new integrated system will continue to automatically create DEM. The ability to change the integrated algorithms that perform a certain function allows the system to change and continue to develop without major re-coding of the framework itself.

Claim 3 *an audit trail and uncertainty information* allows the history and accuracy of a DEM to easily be examined, which promotes a substantiated level of trust in a DEM. These features also allows ready comparison between DEM of the same location by providing information such as which datasets are incorporated, which algorithms were utilized, and output uncertainty.

The following chapters will review the iDEM prototype design in detail. Chapter 2 provides a description of the problem illustrated by the particular use case of tsunami modelling. Chapter 3 introduces the proposed approach to designing a system that tackles the issue of automating DEM creation. Chapter 4 overviews the implementation of the design as a python library, giving representative examples of practical methods utilized in a lightweight partial prototype. Chapter 5 explores fundamental trade-offs in iDEM, and discusses some future considerations of the ongoing development. This discussion of the system covers possible enhancements including providing the library with a user friendly interface to create a fully realized, value added, application. Chapter 6 contains a restatement of the claims and results of the thesis.

Chapter 2

Statement of the Problem and Some Design Considerations

2.1 Problem Summary

A system is required that can produce DEM at various resolutions by combining both topography and bathymetry data while minimizing artifacts and highlighting areas where such artifacts may exist.

The designed system should cope with:

- data from all major elevation measuring technologies;
- a variety of existing data formats;
- data adjustment to eliminate systematic errors;
- some detection and minimization of random errors;
- estimating data uncertainty based on metadata;
- applying appropriate gridding methods that will propagate data uncertainty to build a DCM for every DEM;
- tracking metadata describing:
  - input data characteristics;
  - processing steps; and

   - output DEM characteristics.

- user inexperience by providing a high level of automation; and
- storing all applicable information required to re-create any output DEM.

Such a design will be generally applicable to creating DEM with DCM for most tasks requiring regular DEM. The ability to assess the uncertainty of each DEM created by the system and provide the metadata to track the source data as well as illustrate the DEM and DCM generation process will make iDEM a powerful research tool.

Figure 2.1: Tsunami arrival in Japan (2011) [15].

2.2 A Use Case: DEM for Tsunami Modelling

Accurate tsunami modelling has the potential to save hundreds of thousands of lives. To enable accurate tsunami modelling some methods require multiple nested DEM at various resolutions.

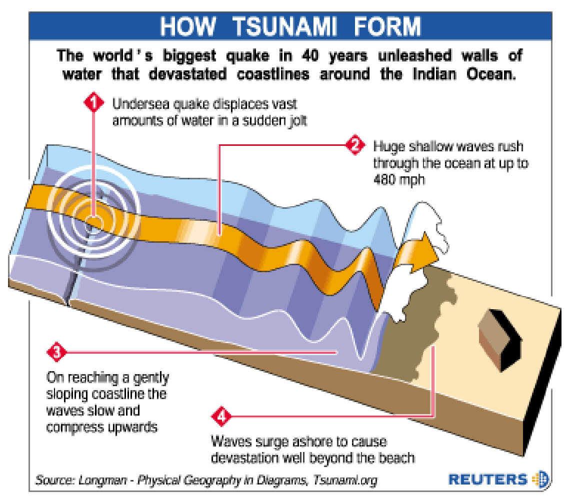


Figure 2.2: Wave elevation increase with shoaling [62].

A tsunami is a series of waves that have a wavelength longer than normal wind generated waves. Tsunami can cause extensive and catastrophic flooding in coastal regions (see Figure 2.1). They can occur in large bodies of water when the water is displaced by earthquakes, underwater volcanic explosions, landslides, glacier calving, meteorite impacts or any other major disturbances. Tsunami can affect an entire ocean basin as illustrated by the 2004 Indian Ocean tsunami caused by an earthquake in Sumatra which affected 14 countries claiming around 230 thousand lives [91, 2].

An typical deep ocean wave shows as a surface variation of about 2 meter in height with a wavelength in the order of 100 meters. A tsunami wave can be as small as 1 meter high with a wavelength of up to 200 kilometres [2]. Such waves can travel as fast as 800 kilometres per hour in deep water but their small amplitude makes them difficult to detect [91, 2].

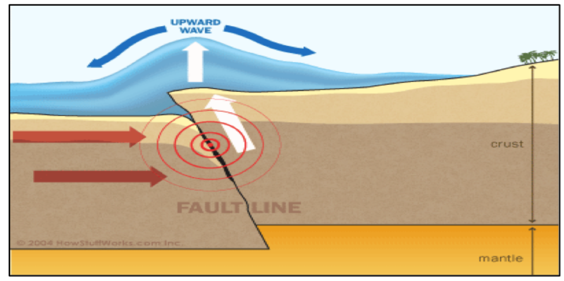


Figure 2.3: Tsunami generation due to an earthquake [67].

When a tsunami approaches a coast, and the water depth diminishes, the wave slows down and the amplitude increases (see Figure 2.2). The speed decreases from hundreds to tens of kilometres per hour forcing the wavelength down proportionally; this results in an enormous increase in wave height. Damage is caused not only by the impact of the high speed water but also by the high volume of water draining off the land, dragging debris with it. Timely warning of impending tsunami, in most cases, allows the evacuation of low lying coastal areas and preparation to minimize damage [91].

The majority of serious tsunami are caused by earthquakes originating in the ocean floor, causing sudden vertical movements at a fault line (see Figure 2.3). The Pacific Ocean is bordered by a ring of fault lines where tectonic plates meet [2].

Tsunami inundation prediction requires representation of the Earth's surface in the form of Digital Elevation Models (see Figure 2.4). To create a DEM, existing digital bathymetric and topographic data is utilized. Such data is generally not extensive enough to provide a simple grid based on the raw elevation points. The existing data points are used as input to calculate a model. Where the data density is higher sub-sampling may be used. Where it is lower an interpolation algorithm can calculate values at the grid points or even extrapolate to areas with little or no data. The resulting DEM is a gridded surface model of the Earth at the resolution of the grid spacing.

Various interpolation algorithms exist (for common examples see 2.7). The surface can be modelled with facets of various shaped polygons; triangles or rectangles are the most common. The facets can be of regular size or irregular size. If irregular facets are used then their size can vary according to data density, the bathymetric depth or surface variability.

Some tsunami models require regular rectangular gridded DEM at various resolutions in a nested configuration. Numerous public and proprietary bathymetry datasets exist, iDEM addresses the need to merge these DEM, using a reliable fusion process, into an overall dataset of the region in order to generate DEM.

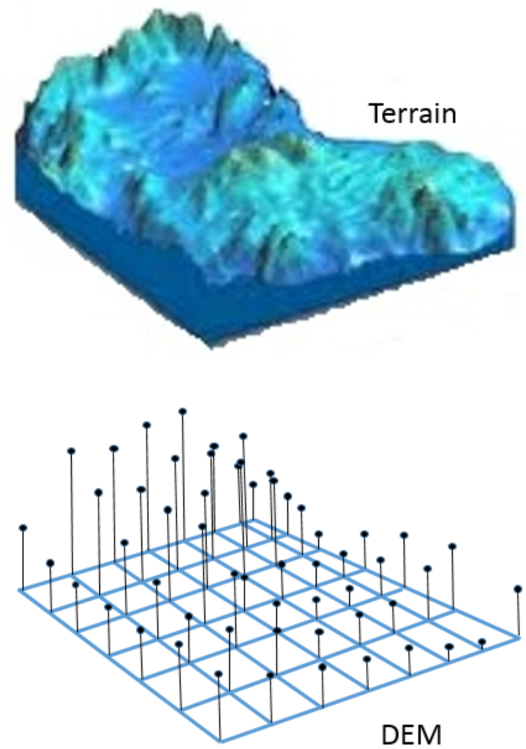


Figure 2.4: Digital Elevation Model (DEM) representing a terrain [71].

In order for researchers, such as tsunami inundation modellers, to be able to directly compare their results they can utilize a sandbox environment containing the DEM. This environment would allow model comparison since all models would be using the same input data. To maximize realism this form of model benchmarking should utilize the best available data.

The proprietary nature of some of the data means that some third party researchers cannot have direct access to this data. Therefore these data cannot be viewed or altered. This restriction will also limit the access to resulting DEM. To give researchers some idea of the quality of the hidden data, they should have access to view the accompanying digital certainty model (DCM). Therefore uncertainty figures will be provided for each data point in each DEM.

The aim of the iDEM research is to design, and then implement, a prototype system to provide the required DEM complete with DCM. The fruits of the iDEM project will be generally applicable and may potentially simplify and standardize DEM creation.

2.3 Data

Extensive work has been undertaken to address the issue of fusing different datasets. When considering elevation data and DEM the solutions assume that the data to be fused is homogeneous in that it has a common coordinate system, common datum, are as free from other systematic errors and are as clean of random errors as possible. In general the starting point of most fusion methods is a collection of homogeneous datasets [17, 81, 48].

2.3.1 Data Collection

One of the reasons that elevation data are not homogeneous is that the systems for collecting such data are varied. On land these data are referred to as topography and are commonly measured from a 'sea level' baseline as a positive value. Ocean depth is referred to as bathymetry and is commonly measured from a 'sea level' baseline as a negative value. The method of collection has a significant bearing on the type and magnitude of both systematic and random errors embedded in the data.

For this research data is not directly collected and existing datasets are utilized as input. A brief synopsis of common methods used in elevation data collection, with a bias to bathymetry (see Figure 2.5), can be found in the rest of this section. The inclusion of this section is to provide an appreciation of some common data collection systems as well as defining terms and acronyms mentioned in later chapters.

2.3.1.1 *In Situ* Sounding

In situ soundings have been carried out from various ocean going platforms such as boats, ships, submarines, remote operated vehicles (ROVs), even robotic autonomous underwater vehicles (AUV) (see Figure 2.6). The tools to measure water depth are as varied as the platforms from which the measurements are made.

2.3.1.1.1 Lead Line Sounding of water depth has been carried out and recorded for centuries. Such soundings were originally carried out by hand using sounding poles in shallow water and a thin rope weighted with a plummet in deeper water. The plummet can be constructed of various materials but was commonly made of lead, which coined the phrase lead line. Such soundings were generally made to the nearest fathom, 6 feet (1.82 meters), and the absolute magnitudes were often rounded

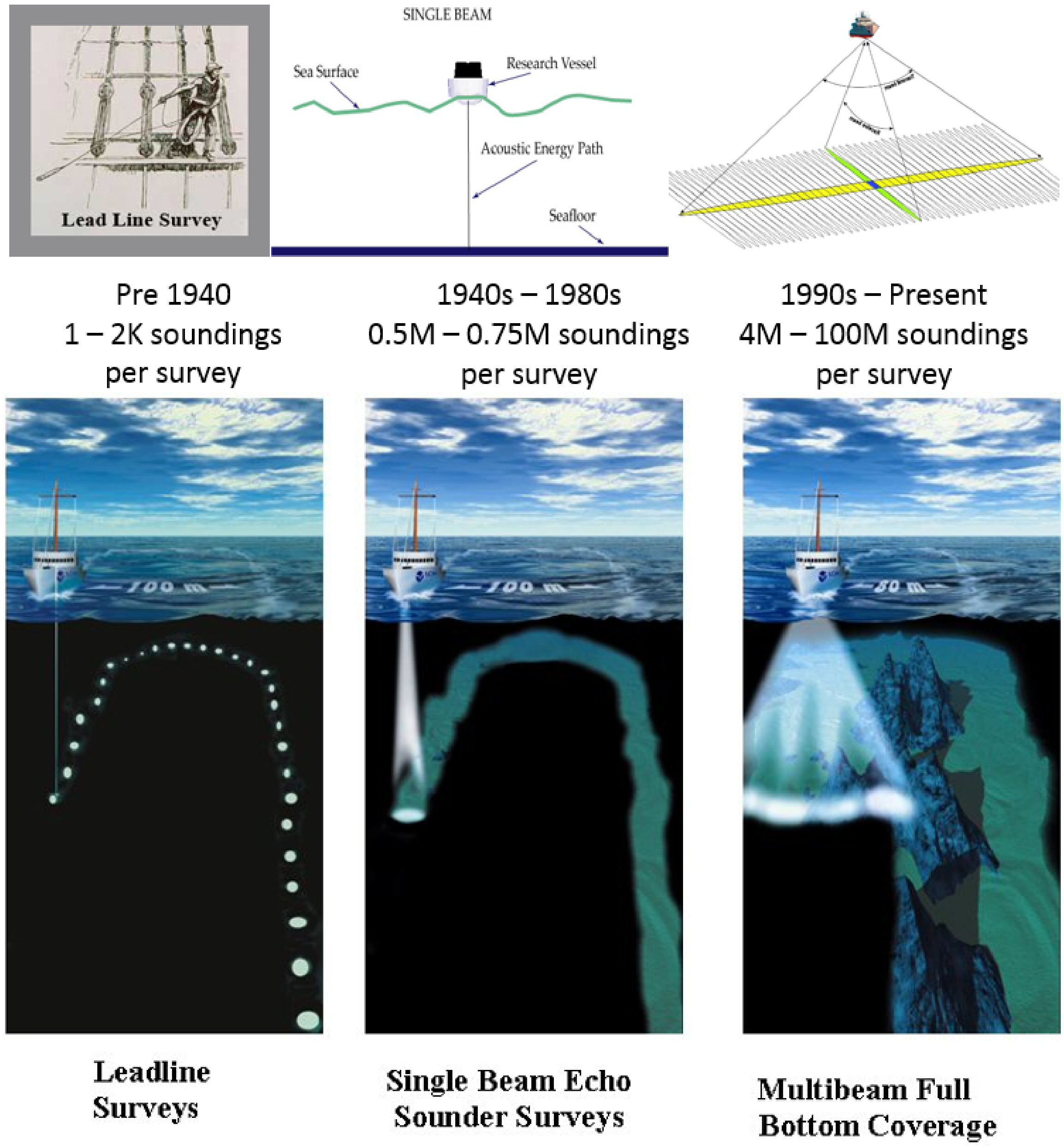


Figure 2.5: Basic comparison of *in situ* sounding technology [92, 26, 69, 65].

down if less than 0.75 of a fathom and up if greater. This rounding biased the results for navigational safety by producing tables and, subsequently, charts that were a little shallower than actually measured. This was a method for trying to ensure that there was always a little more water under a ship's keel, a safety margin when navigating through shallow waters. A more significant consideration is that in the era of lead line soundings the position of a platform's horizontal location upon the surface of the water could be misrepresented by many kilometers in any direction [43, 25, 92].

Figure 2.6: Example platforms for in situ sonar sounding [63, 76, 68].

2.3.1.1.2 Sonar is a technology that uses sound waves propagated through water to navigate, communicate or detect objects on or under water. A passive sonar can be used to detect objects that emit sound, such as ships, whales, dolphins. To measure the depth of the ocean an active sonar is used. An active sonar emits pulses of sound and listens for the returning echoes. Sonar equipment can be calibrated to adjust for various characteristics of the water, such as temperature, salinity, and density. In a properly calibrated system the time it takes for a pulse to return is proportional to the depth [43].

Vertical Beam Echo Sounder (VBES) is also known as the Single Beam Echo Sounder (SBES). First developed in the 1920s, this technology did not supplant lead line sounding until the Second World War. Early VBES systems used a single transducer to ping the bottom (send a sound pulse) and then listen for the return echo. The speed of sound in the water and the time taken for the echo to return is used to calculate the depth. A VBES is effectively a sound line version of the lead line concept. The VBES works faster than a lead line. It gathers more soundings but the system only looks at the depth directly beneath the vessel. Depth is calculated based on the time taken for the sound to echo back to the platform but the speed of sound in water is dependent on varying physical properties such as the temperature of the water. Where lead line data suffered from rounding uncertainty VBES accuracy depends heavily on proper calibration of the speed of sound in the water and thermal profile of the water. Before the advent of global positioning technology VBES positioning accuracy was little better than for lead line sounding [43, 26, 92, 28].

Sidescan Sonar technology grew from the VBES; rather than pinging the ocean floor directly below the vessel the transducer was tilted to one side and a time series of echos captured. A second transducer was added to an array to capture soundings from both sides of the vessel. A minor drawback of this system is that there is no information from directly below the vessel, which is referred to as the nadir. Another minor drawback is that significant terrain features create shadows from which no soundings are returned. The sidescan sonar provides information about the bathymetry to either side of a vessel's track but it is unable to position features accurately or detect the precise depth. This major drawback is due to an assumption being embedded in the technical design concept of the equipment: that the ocean floor is flat [20, 30]. These drawbacks can be minimized by utilizing a VBES system in tandem with a sidescan sonar since their strengths are complimentary [97].

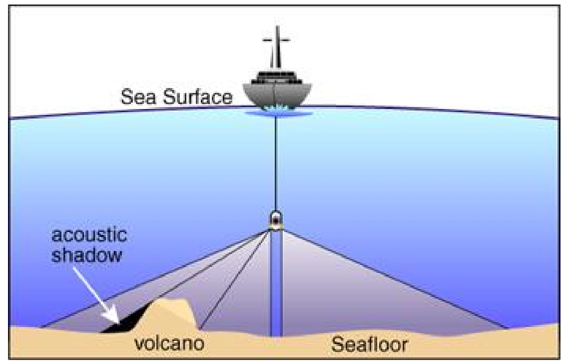


Figure 2.7: Side-scan sonar towfish [74].

The sidescan tow-fish is a sensor platform towed on, or below, the surface of the water (see Figure 2.7). A tow-fish allows greater resolution to be recorded by positioning the platform closer to the ocean floor. Such positioning can also decease the width of the nadir but increases the effect of shadows and makes determination of the position of the tow-fish platform even more difficult. With a Global Positioning System (GPS) the exact location of the tow-fish may be difficult to determine but the location of the towing platform (usually a ship) becomes more exact. An accurate platform position can reduce the tow-fish position uncertainty to as little as ±3 meters depending on the GPS accuracy, the length of the tow line and sea conditions (wind, currents etc) [43, 27, 92, 83, 35].

Sweep Systems are composed of multiple VBES evenly spaced on one or more booms attached off the side of a vessel. This allows for wider coverage of the seabed while still allowing for precise depth and positioning. Some obvious drawbacks are that the booms are awkwardly large, restrict manoeuvrability of the vessel and only allow coverage of the ocean floor to the length of the booms. The uncertainties inherent in the system are similar to those for the regular VBES [43].

MultiBeam EchoSounders (MBES) utilize an array of transducers to transmit pulses and at least one other array to receive the returning echoes. The arrays are positioned to give angular coverage of the ocean floor, and each transducer works on the same principle as the SBES. Unlike the SBES the acoustic beam is very wide in the across-track direction and comparatively narrow in the along-track direction (though still wider than an SBES beam). This allows a wide look angle allowing bottom coverage of about 4 times the depth for most multi-beam echo sounders. Since MBES do not look directly down, booms are not required. In this aspect MBES resembles a sidescan sonar. This system is currently the best technology available. It provides wide across-track coverage at high spatial resolution while overcoming the major issues associated with the SBES, the Sweep System and the Sidescan Sonar. When mounted on a surface platform (such as a ship, see Figure 2.6), used in conjunction with GPS and automatic calibration of the sound speed, the accuracy and breadth of the data collected by this system is high [43, 92, 29].

Phase-measuring Bathymetric Sidescan Sonar looks to the side and captures a time series amplitude of the echoes. The difference between phase measuring and normal sidescan sonar lies in the use of multiple receiver arrays that allow the angel of the returning echo to be determined. This technology provides very high resolution imagery and allows a very wide look angle compared to normal sidescan and MBES. The drawback is a high signal to noise ratio compared to MBES and the existence of the nadir blind region directly below the vessel [43].

2.3.1.2 Utilizing Electronic Charts or Digitizing Paper Charts

For centuries maps and charts have been created representing topography and bathymetry, respectively. The most common form of such representations is the use of contour lines to illustrate the terrain. These contour lines are technically termed isobaths and represent a continuous constant elevation. Digitizing isobaths from existing paper charts, or utilizing those from electronic charts, requires conversion of the analogue contours into discrete elevation points. Many such charts are available. Known issues with these terrain instruments are that the cartographers did not show the actual measured points on a map and that they estimated the position of the contour based on the measured points [45, 10, 34].

2.3.1.3 Remote Sensing

Remote sensing is the act of acquiring information without physically touching the object in question. Both bathymetry and topography can be determined remotely. Topography and shallow water elevations can be measured directly but other bathymetry needs to be inferred through gravity modelling using Radar Altimetry. The two most common platforms for such remote sensing are aircraft and satellites.

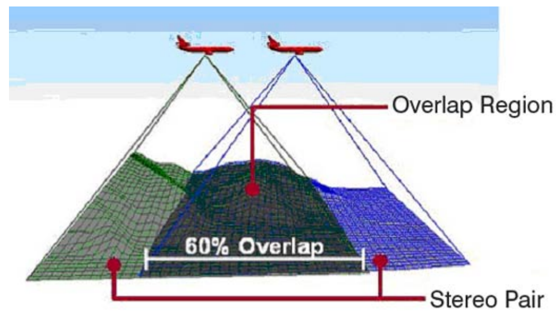


Figure 2.8: Stereophotogrammetry [72].

2.3.1.3.1 Photogrammetry is the remote measurement of exact positions of surface points from photographs. Topographic DEM can be created using pairs of aerial photographs taken from different positions (see Figure 2.8). Accuracy estimation is not inherent in the photogrammetry process.

2.3.1.3.2 LiDAR uses laser pulses to determine distance (see Figure 2.9). A red, or infrared, laser system has few limitations for topography although processing is required to eliminate ground cover such as trees. This colour of laser light is not appropriate for bathymetric measurements. The red laser is used to determine the height above the water while a blue-green laser is used to measure depth because it provides a return signal from the ocean floor. A blue-green laser produces a reasonable return for water depths of up to 50 meters when the water is clear and the ocean floor is a light color. Noise can be generated by water

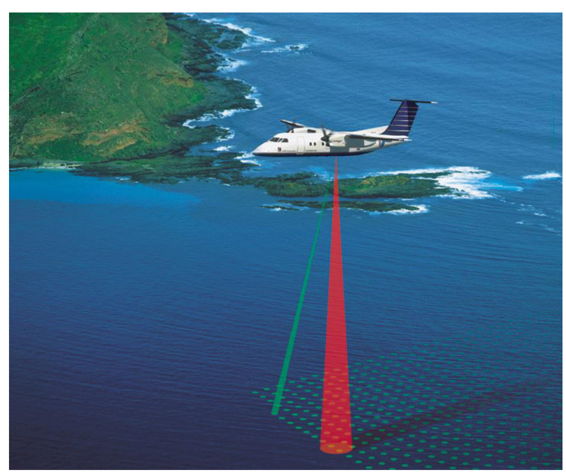


Figure 2.9: Aircraft LiDAR [73].

that is turbid or subject to suspended material such as air bubbles, fish or seaweed [44, 96, 101]. Lidar produces high density data in areas that sonar platforms find difficult to measure. The noise translates into higher uncertainty and erroneous data outliers.

2.3.1.3.3 Radar Altimetry is used to directly measure topography elevation. The radio waves are sent down to the ground and the time taken for the echo to return to the satellite is used to calculate the elevation of the terrain (see Figure 2.10). Water rapidly attenuates the radio waves so this method can not be used to measure bathymetry directly but the water surface does provide a return signal. If the ocean floor is perfectly flat and the wind calm then the surface of the ocean is flat. If an underwater mountain exists on the flat ocean floor that mountain will generate a tiny amount of extra gravity at its location. The extra gravity exerts a little pull on the water creating a slight mound over the location of the mountain. The water level can be measured with great accuracy using radar altimetry. Gravity modelling can use such measurements to estimate the bathymetry beneath the water [117].

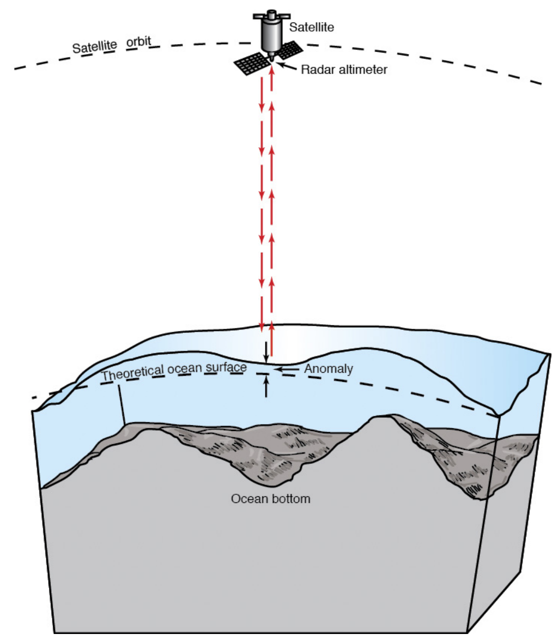


Figure 2.10: Radar Altimetry [75].

2.3.2 Data Formats

Most measuring systems use different hardware designs, different processing schema and different storage formats. These differences affect the accuracy of measured information as well as the type and quantity of relevant meta-data. In some cases the elevation and/or position information is not measured directly but is instead calculated or modelled based on some other measured attribute. All these technologies use different formats for storing the data. Applications have been developed to view, edit, manipulate and analyse the data. The more versatile of those applications read many different formats, often translating the data into the native format of the application.

Two distinct format categories have emerged, one based on representing each elevation data point as a vector and another based on raster representation [33]. Elevation information is often visualized as contour lines, a common version of the vector category format is to represent such terrain as data lines rather than data points. The contours are defined by sets of points joined in relationships that indicate how lines are to connect the points to create the contours. Such vector information is stored in shape files where each contour is an object, or shape, defined in the file. By definition a DEM is a collection of elevation points, the points are not related to each other in ways that defines objects. DEM vector files are commonly referred to as XYZ files, these store position and elevation information for each data point. It is possible to convert a shape file into a point file but processing is required for the reverse operation. A shape file generally contains less information that the DEM from which it was processed because any data points between the contour lines are discarded. The raster format assumes that the elevation data is stored in a regular grid layout. A common raster file format stores only one position as a reference point and then it stores the x and y size of a raster cell (e.g. pixel) and the total number of x cells and y cells. With this information an application can calculate the position of each and every raster cell in a raster grid. The raster file then stores a list of elevation values corresponding to the position calculated for each raster cell. There are variations on this theme that store slightly different information but the general principal of the raster file format is the same [8, 116, 98, 82, 103].

Raster files have the drawback of needing to store a 'no-data' value for each raster cell without an elevation value. Despite this drawback a raster file is generally significantly smaller than a corresponding vector file because it does not store detailed position information. Optimally, a fully populated rectangular XYZ DEM can be represented as a raster file that approaches 1/3 the size of a vector file.

Both vector and raster categories have many different implementations of their respective styles. Specialized applications have been developed just to handle the translation between the many formats. Two such common applications are GDAL, which handles in excess of 130 raster formats, and OGR, in excess of 80 vector formats. These two application are routinely embedded in other applications as the core translation functions [49].

The International Hydrographic Organization is working to introduce a new standard data object. The Navigation Surface data object represents bathymetry, uncertainty, and methods for the manipulation, combination and utilization of such objects. The Bathymetric Attributed Grid (BAG) file format stores elevation, uncertainty, and metadata in the same raster file. The importance of the format is that it is designed as a universal standard that can contain raw survey data through to certified navigation DEM in the same file format. The uniform file format is designed to allow data to be passed seamlessly between different software applications and organizations [11, 123]. Many existing applications that handle DEM are incorporating this standard as an optional format [49, 41, 11].

There would appear to be some distinct advantages to the BAG data format. This format ensures that minimal metadata and available uncertainty information accompanies the elevation information. In most other formats the metadata and uncertainty information are stored in separate files. Many formats store limited metadata with the elevation information, some formats allow uncertainty information to be stored in the same file. In many instances, with separate files, detailed metadata is lacking or, as with the uncertainty, is missing entirely. Regardless of which particular format is used to store such data issues still arise. Sometimes these issues arise from how a format has been utilized, in other instances they occur because the metadata is inaccurate; or the data itself is other than expected.

The occurrence of missing metadata and uncertainty information are very common. Inaccurate metadata is another matter which can be simpler to correct though more difficult to diagnose. During this research an inaccurate dataset was encountered that illustrates the effect of such inaccuracies.

BAG raster file that is populated with elevation, metadata and uncertainty information. Each type of information can be seen to populate a different layer of the data file. The elevation and uncertainty layers can both be visualized as arrays of values where each value represents the elevation or uncertainty at a particular location. For example, metadata describes characteristics about the elevation and uncertainty information such as vertical datum, horizontal datum, grid size, etc.. In the metadata is a variable that defines a 'nodata' value for elevation and another for uncertainty. These 'nodata' values are inserted into the elevation and uncertainty arrays where elevation or uncertainty information is lacking. If the 'nodata' value is defined as zero for elevation then a value of zero is inserted at each point in the elevation array for which an elevation value is unavailable. When the file is read these points can be ignored or processed in some particular fashion according to need. Problems occur when the ’nodata’ values are wrong. If a ’nodata’ value is set to -32,000 for uncertainty and the uncertainty array is populated only with -32,000 values then it will correctly appear as though no uncertainty data exists. If the array is accidently populated with zeros then it will appear that the elevation data is perfect with zero uncertainty. It is possible to correct such an error in two ways, an application can open the file and reset the uncertainty ’nodata’ value or an application can be told that such an error exists and to treat the zeros as ’nodata’ despite the metadata. In both cases human intervention is required to detect the error and decide on a corrective strategy.

Another encounter illustrates an issue with repurposing data. A file format was encountered that incorporates elevation and some metadata. This was not a BAG format but had similar functionality. The metadata was correct, as was the elevation data but the purpose of the data was topographic intending to show land and ocean surface. In a global DEM the ocean floor, not its surface, is usually required. In order to utilize such data it needs to be filtered to remove all ocean surface values. In this case such values were zero so it is possible to simply filter out all zero values but such a filter would also remove valid coastline information. A better filter should be used that removes all water surface values except along the coast where they define the coastline. Again it is up to the user to be familiar with the data so that it can either be preprocessed or so that the application can be instructed to deal with the issue by applying the appropriate filter.

2.3.3 Data Sources

Many different organisations gather raw elevation data and create DEM. Both topographic and bathymetric data is generally available from some of these organisations. Other data are considered proprietary; some can be purchased, while certain data is unavailable, being considered sensitive. The following is a synopsis of some sources that can provide elevation data:

- Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) has been utilized to create a topographic DEM of the globe. This DEM is a photogrammetric product created from stereo images obtained by the ASTER instrument aboard the Terra satellite [119].

- General Bathymetric Chart of the Ocean (GEBCO) is a publicly available global DEM. This DEM has been compiled from ship soundings with bathymetric interpolation referenced to gravity data while land topography is predominantly based on SRTM30 altimetry [50].

- National Oceanic and Atmospheric Administration (NOAA) is a United States government organisation that falls under their public availability rules for government funded data. Bathymetric data held by NOAA is available for download on-line at no cost. Most of the data available off the coast of British Columbia were collected using SBES in comprehensive low resolution surveys. Extensive MBES data is primarily available along main shipping routes. A small selection of BAG format with uncertainty is available for selected US coastal areas. The data is available in raw format as sounding datasets or as processed data in the form of DEM at a variety of different resolutions [95].

- Natural Resources Canada makes available to the public, topography on predefined or custom defined geographic areas of Canada [13]. An API is available for developers to directly access elevation models from other applications.

- Shuttle Radar Topography Mission (SRTM). SRTM30 Plus V11 is the latest topographic and bathymetric gravity interpolation DEM of the globe. The bathymetric gravity interpolation is validated with existing sounding data [99, 5, 119].

Information is also available from less accessible sources. Some companies sell information that they have collected, others share the information under licence or by agreement. Table 2.1 on page 23 contains example sources and lists a few factors that highlight some differences between data from different sources.

	ASTER	GEBCO	GEBCO	NOAA	NRC	SRTM	SRTM
Formats							
ARC				X			
ASCII				X			
Binary Float				X			
Esri ASCII raster		X					
GeoTiff	X				X		
GeoTiff INT16		X					
netCDF		X	X	X		X	
USGS HGT						X	
XYZ				X			
Projection							
Geographic (4326)	X	X	X	X		X	
Geographic (4617)					X		
Lambert (3979)					X		
Data type							
Topographic	X	X	X	X	X	X	X
Bathymetric		X	X	X		X	
Resolution							
1 arc minute			X	X			
30 arc seconds		X		X		X	
15 arc seconds				X		X	
12 arc seconds					X		
6 arc seconds				X	X		
3 arc seconds				X	X		X
1 arc second				X			X
3/4 arc second					X		
30 meters	X						
other				X			
Uncertainty (average)							
absolute 95% meters	17.1			-		7.86	
std deviation meters	8.68			-		3.95	

Table 2.1: Data sources: Illustrating the variety of formats and characteristics.

2.4 Systematic Errors

Data points from different datasets need to be adjusted to use the same coordinate system and to have a common horizontal and vertical datum. Systematic error sources can be eliminated by adjusting frames of reference [17].

2.4.1 Coordinate Systems

Local coordinate systems are often utilized to project internally consistent, scale models or maps. The global coordinate systems are usually longitude and latitude based, producing the flat maps of the globe with which we are most familiar. Such maps are not internally consistent in that the higher latitudes depict the terrain in an increasingly distorted manner. This can best be seen in countries closer to the poles, such as Canada, which appear large but are in fact smaller than they seem in comparison to countries on the equator. Various projection systems have been devised to reduce these distortions [87].

2.4.1.1 Universal Transverse Mercator (UTM)

This coordinate system uses a metric Cartesian grid on local surfaces with conformal projection (see Figure 2.11). The projections are local because a specific set of UTM projection parameters are only valid within a defined area of the globe. Each such area is termed a zone. In order to create a metric Cartesian grid outside the local zone a different set of UTM projection parameters need to be utilized. This means that UTM is a series of local map projections. The projections do not include the polar regions, which are covered by the Universal Polar Stereographic (UPS) system. Each projection in the UTM system covers a 6 degree band of longitude, resulting in 60 separate projection zones. This system produces terrain models that are internally consistent, properly proportioned and to scale.

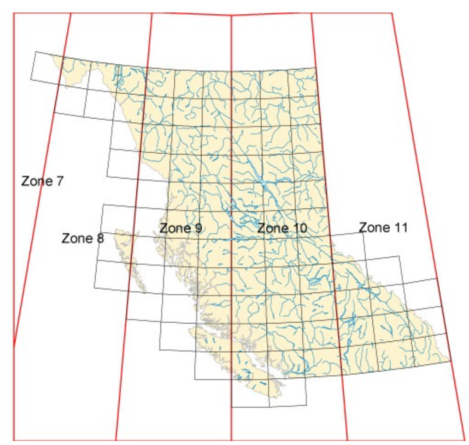


Figure 2.11: UTM zones, BC Canada [64].

2.4.1.2 Geographic Latitude and Longitude

Geographic latitude and longitude are measured in degrees. Latitude of 0 degrees is anywhere on the equator, 90 degrees (north) is the north pole and -90 degrees (south) is the south pole. Longitude is measured around the equator where 0 degrees is a line stretching from the north pole to the south pole passing through the Royal Observatory in Greenwich, United Kingdom (see Figure 2.12).

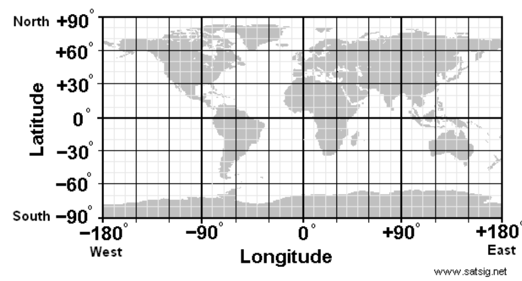


Figure 2.12: Geographic latitude and longitude [66].

2.4.2 Position Datum

A *datum* is a fixed starting point, a reference. Various counties and organizations around the world employ many different datum to define reference points for position and elevation of their bathymetry and topography data. A datum together with the coordinates of the point give you the position for that elevation value. The same position information referenced to a different horizontal datum will redefine the position to another part of the world. The same datum method applies to elevation information. The vertical datum reference used defines the exact elevation at a particular position. Using the wrong vertical elevation datum at a particular point can make a significant difference to the apparent elevation at the point.

The next four sections will review the most relevant datums for this project.

2.4.3 North American (horizontal) Datum: NAD27, NAD83

The North American Datum of 1927 (NAD 27) was the horizontal control datum for the United States. The Earth's surface is approximated by an ellipsoid which is a smooth slightly distorted sphere. NAD 27 was defined by the Clarke ellipsoid which was tailored specifically to fit North America. The origin of the Clarke ellipsoid is referenced to the Meades Ranch survey station with the assumption that the station's geoidal height was zero. There are many ellipsoids in use by different countries and each one aims to minimizes the deviation between the ellipsoid and the local terrain.

The Clarke ellipsoid was defined by fitting an algorithm to position measurements using optical levelling survey data. The location of specific control points across the continent were all measured from, or referenced to, the Meades Ranch survey station as the origin. NAD 27 was the horizontal datum determining the longitude and latitude coordinates of every point on the continent.

A *geoid* is the hypothetical shape of the earth, coinciding with mean sea level and its imagined extension under (or over) land areas. As technology and needs advanced it was recognized that the Clarke ellipsoid of 1866 could not serve the needs of a modern geodetic network. Using over 250,000 reference points including 600 Doppler satellite positions a new ellipsoid was computed. This ellipsoid is based on a *geocentric* (center of the Earth) origin and removed significant local horizontal position distortions.

NAD83 remains the official horizontal datum for the United States, Canada, Mexico, and Central America [14, 102, 94, 17].

2.4.3.1 North American Vertical Datum: NAVD27, NAVD88

The previous horizontal map datum was North American Vertical Datum 1927 (NAVD27). It was originally named the Sea Level Datum of 1929 because it was based on measurements of mean sea level at 26 tide gauges scattered through the United States and Canada. The resulting datum does not conform to mean sea level, or any equipotential surface, so it was renamed in 1973 to NAVD27.

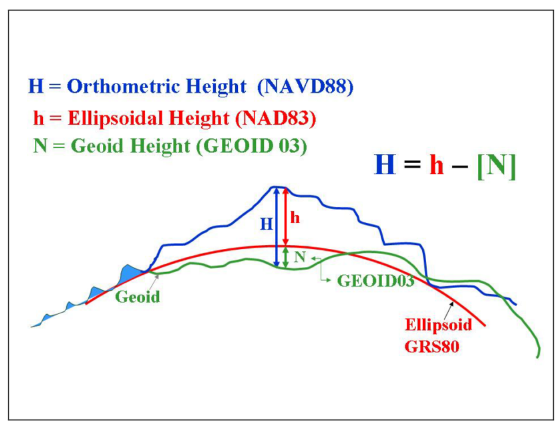


Figure 2.13: The relationship between NAD83 and NAVD88 [120].

The shape that the surface of the oceans would take under the influence of Earth's gravitation and rotation is called the geoid. Wind and tide influence is not part of the geoid. Height measured from a geoid conforming to mean sea level is termed *orthometric* height.

The North American Vertical Datum 1988 (NAVD88) is an updated version of NAVD27 (see Figure 2.13)[94]. In the United States the current standard for orthometric height reference is the North American Vertical Datum NAVD88 which is referenced to the NAD83 ellipsoid (GRS80). NAVD88 is more accurate than NAVD27, and neither datum correlates fully to an equipotential sea level. Both datum are based on reference point measurements relative to a defined origin, geodetic levelling. This levelling was carried out using optical levelling survey data. The height at specific control locations across the continent were all measured from, or referenced to, the Meades Ranch survey station as the origin. Using this method requires further reference point measurements to maintain accuracy over time because the Earth's crust moves over a relative short time frame [17, 94].

2.4.3.2 Canadian Vertical Datum: CGVD28, CGVD2013

The Canadian Geodetic Vertical Datum of 1928 (CGVD28) is a geodetic datum that defines orthometric height. Figure 2.14 shows how CGVD28 and NAVD88 relate to terrain and sea levels.

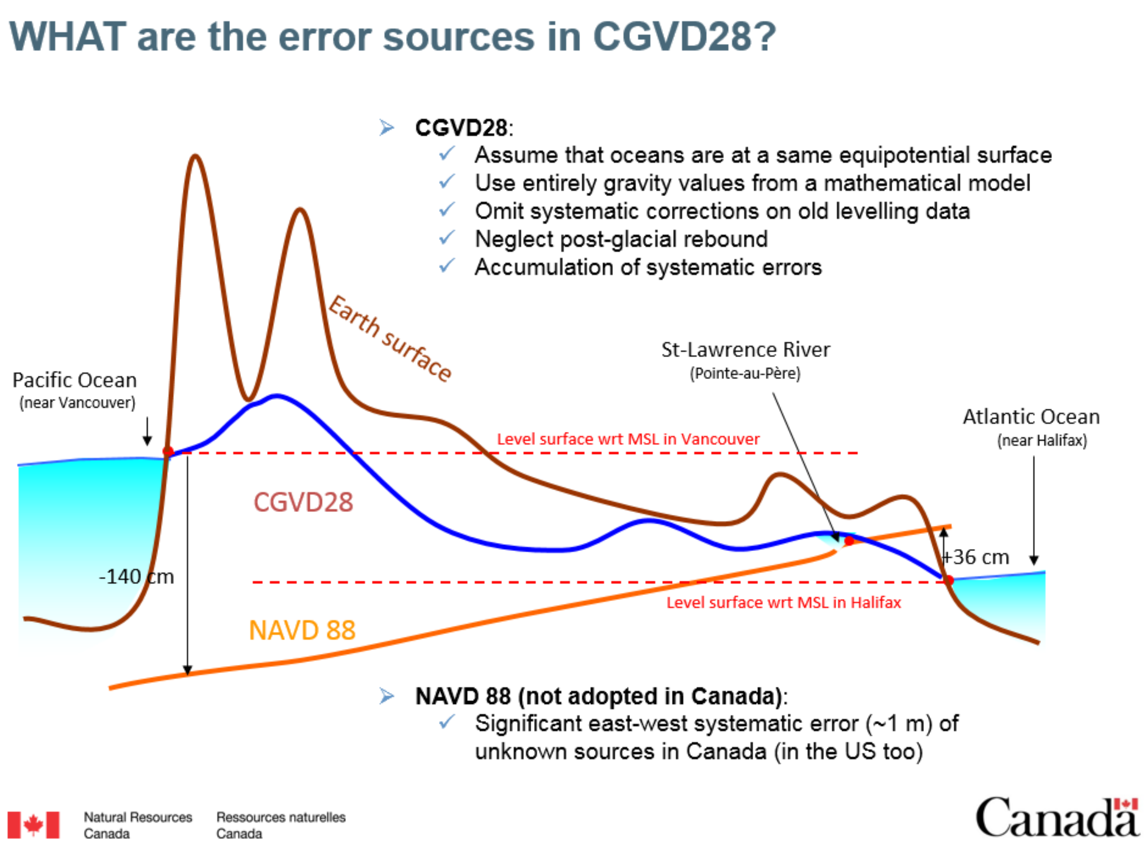


Figure 2.14: The relationship between CGVD28 and NAVD88 [120].

As with NAVD27 and NAVD88 the CGVD28 geoid is based on geodetic levelling. This datum was superseded by Canadian Gravimetric Geoid Model in 2013 (CGGM2013). CGGM2013 is also known as Canadian Gravimetric Vertical Datum 2013 (CGVD2013).

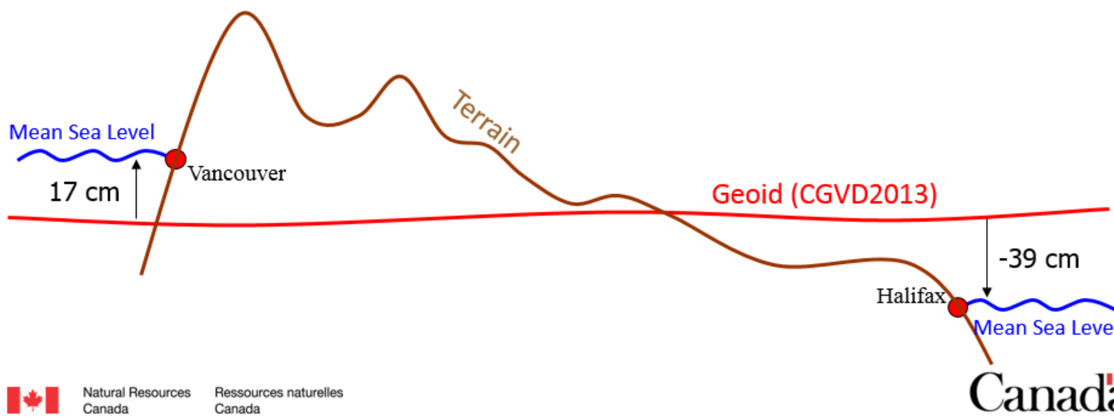


Figure 2.15: CGGM2013 and relative sea levels [120].

CGGM2013 is not based on geodetic levelling. It is a gravimetric geoid model referenced to the horizontal datum ellipsoid GRS80 (see Figure 2.15). Unlike the Earth's crust the gravity field can be considered static so the geodic model will not change appreciably over time. The accuracy of the model will continue to be improved. This new datum is being phased in over time in Canada. The United States, Mexico, Central America, Caribbean and Greenland plan to adopt similar models by 2022, creating a seamless vertical datum for the North American continent[120].

2.4.3.3 Tidal Datum

Tidal datum are vertical datum (see Figure 2.16). These datum usually only apply when areas that include coastal waters are of interest. The uncertainty of bathymetric elevation measurements has a component that is proportional to water depth. In deeper waters the uncertainty of the measurement is greater than the tidal variation so the tide can safely be ignored since the tidal variation is within the uncertainty bounds of the depth measurement. If the area of interest is topographic then tide is usually not a factor as a tidal datum would not normally be chosen.

For a clearer understanding of this it can be considered that there is a distinction between a tidal datum and a more general global vertical datum. A general vertical datum can be defined as a spheroid (see Section 2.4.3.1 & 2.4.3.2) that can encompasses the whole globe and thereby define a zero height surface that includes a nominal sea level. Topography can be referenced directly to such a vertical datum. The nominal sea level does not change with time. Bathymetry is also referenced to a vertical datum but it is not necessarily a temporally static one, as tides cycle over time. The changing tides effect the measurements of bathymetry. Water level will also effect the outcome of events; for example:

- the tidal level at a particular time determines the clearance for the passage of water craft; and
- the height of the inundation of a tsunami will be greater during high tide than during low tide.

To overcome the temporal nature of the tide the changing water levels are averaged over a long period of time, usually 19 years. Various tide levels are defined as tidal datum. There are usually two high and two low water in a given tidal cycle; one while the moon is on the near side of the planet and one when it is on the far side. The water levels are usually different for each cycle. As a result there are many different ways to define a tide level; for example, MHHW is the average of a collection of the higher high water levels in each cycle.

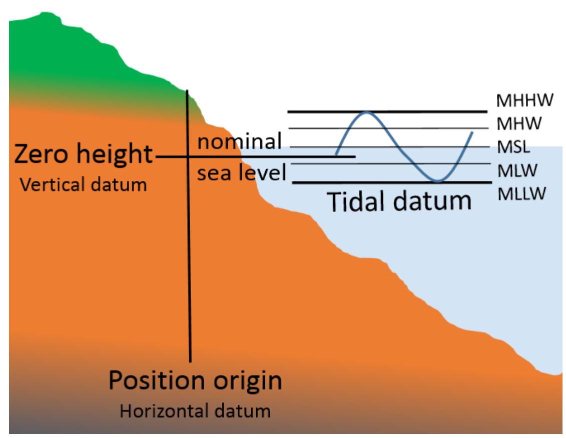


Figure 2.16: Datum

Some of the more common tidal sea datum are [24]:

- mean higher high water (MHHW) - average of the elevations of the higher high water of a mixed tide;
- mean high water (MHW) - average elevation of all high waters recorded at a particular location;
- mean tide level (MTL) - halfway between mean high water and mean low water;

- mean sea level (MSL) - average height of the sea for all tides;
- mean low water (MLW) - average height of the low waters; and
- mean lower low water (MLLW) - average height of the lower low waters.

The following terms are used to define the tide levels:

- diurnal - of the tidal day.
- tidal day = lunar day which is approximately 24.84 hours, the time the moon takes to rotate around the Earth, the interval between two successive upper transits of the moon over a location.
- mixed tide - the presence of a diurnal wave is conspicuous by a large inequality in either the high or low water heights with two high waters and two low waters usually occurring each tidal day.

These defined tidal datum are not general in nature, as the tides are often measured and datum defined locally. The result is that the height of Mean Sea Level datum in one coastal region is not the same as that in another region. The difference between these local mean sea levels and the general nominal sea level datum are unique for each local Mean Sea Level datum. Models are being created that use the averaged local tidal values to create a set of defined tidal datum surfaces. At present such models may cover part of a coastal region of a continent. In the future a complex model may exist that complements the global definition of nominal sea level with a set of surfaces that accurately define the various tide levels, allowing simple tidal datum adjustment regardless of location.

Vertical datum adjustment of topography is somewhat more complex in coastal regions due to the localized tidal datum. For example, consider Vancouver Island in British Columbia (Canada); the Mean High Water level is different on the west coast as compared to the east coast. If a combined tidal datum model exists that encompasses both west and east coasts, while also defining the tide level beneath the island, it is (for example) simple to adjust the bathymetric and topographic elevation values from the nominal sea level datum to the Mean High Water level. This can be done by the simple application of an adjustment of the difference of the two datum surfaces to the elevation values. If such a model does not exist then the local Mean High Water adjustment for the west coast and the east coast can be used to estimate the required adjustment.

There are, in fact, three main options available for tidal datum adjustment in waters off the coast of British Columbia:

- estimation using: [17]:
  - the Harmonic Constant Datum Method (HCD). An efficient way of estimating tidal datums relative to mean sea level. Does not require a long time series. Is prone to discontinuities between mixed and diurnal tidal regimes.
  - the Interpolation Method. A tidal benchmark is a combination of tidal gauge observations and the gauge's relationship to the geoid. One benchmark can be used to represent tidal datums at the benchmark location. If multiple benchmarks are available, at various locations, a tidal surface can be created by interpolating the information of all the gauge reference stations.
- applying the two existing current tidal datum surfaces (together they give complete coverage of the B.C. coastal region) [36, 47].

It should also be noted that the new Canadian Continuous Vertical Datum for Canadian Waters (CVDCW) is in prototype testing at present. This tidal datum is a versatile model covering all Canadian coastal waters. The CVSCW references vertical datum CGGM2013 and will replace existing datum for Canadian waters. In 2014 the CVDCW was available for the Arctic and Hudson Bay. The CVDCW should evolve quickly in the next few years with prototypes being validated against new GPS and gauge observations. The completion of this model will allow accurate vertical datum adjustment to any tidal level for all Canadian coastal waters [110].

2.4.4 Global Datum: WGS-84

The World Geodetic System (WGS) is actually a standard that incorporates both horizontal and vertical datum that are applicable for the whole globe. WGS-84 is a version of the WGS global datum that was substantially revised in 1984. Minor revisions have occurred regularly since then, and no revisions have been required since August 27, 2007. WGS-84 continues to be used as a major global datum.

WGS-84 incorporates:

- standard global coordinate system, geographic longitude and latitude,
- horizontal datum referenced to the Royal Observatory in Greenwich United Kingdom,
- vertical datum EGM96 (Earth Gravitational Model 1996), a geoid defining zero height including nominal sea level.

The EGM96 vertical datum defines topography as positive measurements of elevation relative to the geoid and bathymetry as negative elevation measurements (depth) relative to nominal sea level. This 3D positioning system is used by the GPS satellite navigation system and for NATO military geodetic surveying [108, 1, 4].

Using WGS-84 as the default datum will allow data from anywhere on the globe to be stored and processed. When exporting from the proposed system, stored data and DEM can be transformed to local datum as desired.

2.5 Random Errors

2.5.1 Outliers

An *outlier* is a data point that lies abnormally far from the majority of other data points in a data set. When discussing DEM, an outlier is usually caused by a measurement error. For example, bathymetry errors can result from intermittent equipment glitches, sea conditions, water turbidity or the presence of obstacles interfering with the process. Obstacles can be fish, bubbles, vegetation or debris. It is also possible that the terrain itself affects the measurement: deep low density sediment can result in higher than actual readings. Such data errors occur regularly and can have a significant effect on a DEM (Austin, 2009). The outliers create non-existent surface features that are commonly termed artifacts. Automatic outlier detection and editing is a difficult task. Recent work in this area tends to concentrate on automatic processing of MBES data [59, 85, 9]. One culmination of such research is central to a leading MBES processing application: Combined Uncertainty and Bathymetric

Estimator (CUBE). Outlier algorithms of a more general nature can be applied to other elevation data [46]. The CUBE application has notable success with automating outlier preprocessing but expert user intervention is still recommended, and required for total artifact elimination [18, 19].

2.5.2 Sampling Distortion

A lack, or sparsity, of data can only be solved by incorporating additional datasets with samples of the sparsely populated region. On the other hand an over abundance of data can be adjusted.

Aliasing refers to inability to reconstruct a signal from samples; such as a terrain surface from a DEM. The DEM is based on samples interpolated from elevation measurements. If more signal points (elevation measurements) exist than grid nodes (sample rate) then the signal includes high frequency components that cannot be

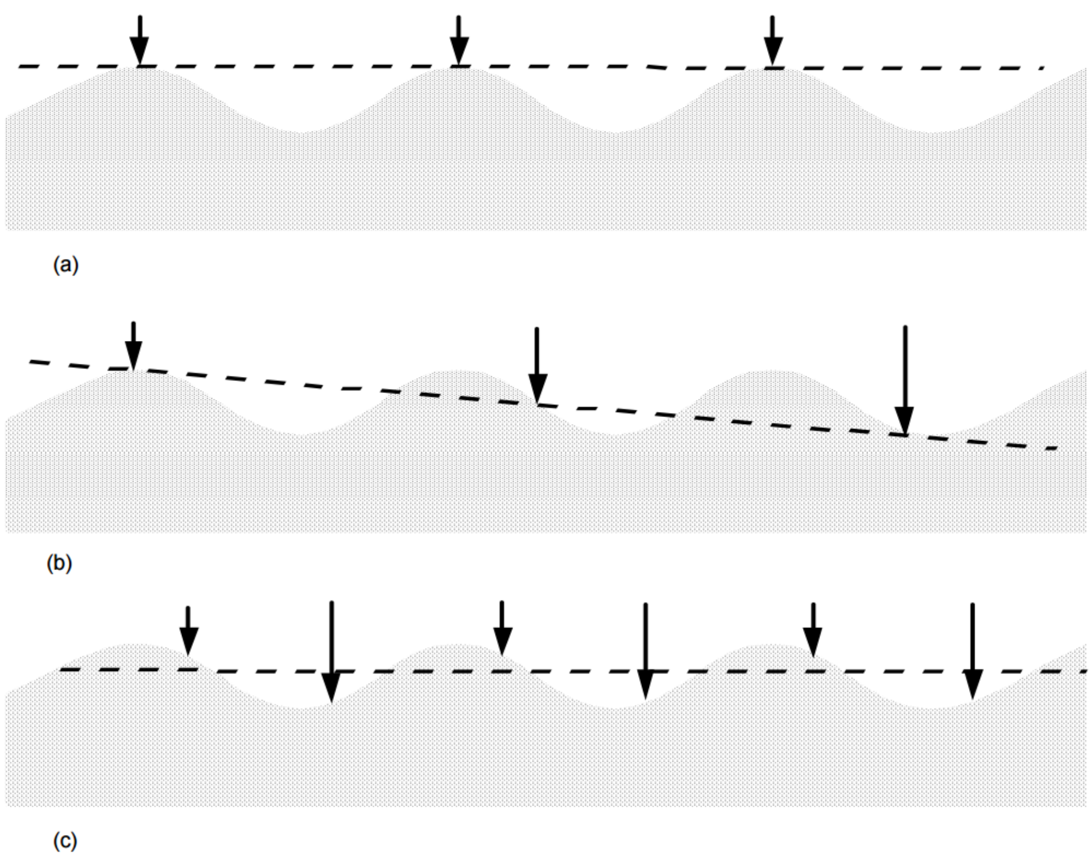


Figure 2.17: Example of aliasing [52].

accurately reproduced by the DEM. In DEM the sample frequency is equivalent to the reciprocal of the grid cell size $1/(grid\ size)$, equivalent to the Nyquist frequency. For a surface component frequency to be correctly reconstructed by the DEM if must be sampled at twice its rate. The maximum frequency that a DEM can accurately reconstruct is half the Nyquist frequency. Since the sample rate, Nyquist frequency, is equivalent to the grid size that means that the maximum surface component frequency that can be represented accurately is equivalent to $1/(2 \times grid\ size)$. In effect a grid cannot represent variations smaller than the grid size therefore rather than attempting to grid multiple values in a grid cell they should be aggregated to give the average depth in an area defined by the grid cell $(grid\ size)^2$. Such grid block aggregation can be used as a crude low-pass filter to reduce aliasing in surface interpolation [121, 118, 52].

2.6 Uncertainty

Processing elevation data to create DEMs can also create related uncertainty surfaces. Some of the data has associated uncertainty figures for horizontal positioning and/or elevation. This uncertainty can become part of the input for an appropriate interpolation process yielding both a DEM and a DCM. To create a DCM the elevation uncertainty is a requirement

Uncertainty is derived from many sources in each of the systems used to gather elevation data. These uncertainties propagate through the system, are usually increased by the processing, and then combined to give rise to a total uncertainty for the measurement. The final uncertainty is referred to as the Total Propagated Uncertainty (TPU). The TPU can be represented as variance ($uncertainty^2$) with units of $meters^2$, but is usually reported as uncertainty in $meters$, with an associated statistical confidence level. These confidence levels are derived from the assumption that the uncertainty follows a Gaussian distribution. It is also commonly assumed that the horizontal position (X, Y) uncertainty is *isotropic*, the same in every direction, which results in one value of uncertainty that relates to both the X and the Y coordinates. For bathymetry the elevation (Z) uncertainty usually varies in proportion to the water depth through which the elevation is measured. With topography the vertical elevation (Z) uncertainty is derived from various fixed uncertainties where these fixed uncertainties do not vary with the value of elevation being measured [109]. Most bathymetric uncertainty is compiled using the International Organization for Standardization (ISO) quadratic summation method [61]. This method is not conservative like an arithmetic summation, the resulting total propagated quadratic summation uncertainty is smaller that the simple arithmetic summation would be. The following example is given in IHOs S-44 Standards for Hydrographic Surveys:

$$TPU = \pm\sqrt{a^2 + (b \times depth)^2} \tag{2.1}$$

a: is a fixed uncertainty b: is a depth variable uncertainty

The equivalent topographic equation is [55]:

$$TPU = \pm\sqrt{a^2 + b^2} \tag{2.2}$$

a: is a fixed uncertainty b: is another fixed uncertainty

2.6.1 Uncertainty Estimation Methods

A significant portion of the available elevation data does not have specific uncertainty data. There are methods that can estimate the uncertainty based on preliminary estimates of accuracy of the horizontal positioning and vertical elevation. Such estimates are based on the technology used to collect the data or, if such metadata is unavailable, on the era in which it was collected. For higher density data such as MBES the Combined Uncertainty and Bathymetric Estimator (CUBE) produces notable results. For data that cannot be processed by CUBE the simplest solution employs Monte Carlo processing [79] in the following way:

1. The dataset points are selected by the system. Each point is assigned an initial estimate of horizontal position uncertainty (U_h).

2. The data point positions are varied by a random normal distribution of (U_h), (longitude + $random(U_h)$, latitude + $random(U_h)$).

3. The randomized dataset is gridded and interpolated to create a DEM with desired grid size (resolution).

4. The elevation values of DEM at the actual positions of input data points are determined.

5. The difference between the original point elevation value and the calculated Monte Carlo point elevation value is determined and saved for each point.

6. Steps 2 through 5 are repeated 100 times.

7. For each point, analyse the 100 saved difference values to estimate its uncertainty.

Monte Carlo calculations are computationally intensive, which makes uncertainty estimation of large datasets, especially large high density datasets, slow. A Bayesian method exists that utilizes Monte Carlo sample processing could be more effective in some cases and is significantly faster [39]. A generalized method based on the CUBE approach to uncertainty estimation is also possible. This method is not computationally intensive [38].

Initial uncertainty information for horizontal positioning and elevation is required to apply Monte Carlo estimation. It is possible to assign initial uncertainty values inferred from the era of the data collection, the platform used, the method used, and the location surveyed when uncertainty information is not provided in the meta-data.

The iDEM system design incorporates the Monte Carlo algorithm as the initial estimation method. Other methods can be added in the future based on better or more efficient algorithms.

2.6.2 Sources for Inferring Initial Uncertainty

Detailed knowledge of the various uncertainties embedded in an elevation measuring system are not required since the elevation uncertainty can be estimated from the position uncertainty. Position uncertainty is dependent on the technology used to determine the position of the measuring platform.

Often the exact information pertaining to initial uncertainty estimation is not available. Under many conditions it is safest to assume a worst-case scenario [80]. Determining applicable worst case scenario figures may require some work. In an automated system a starting point could be to utilize the largest uncertainty values documented for similar measurement situation using similar technology. Some documented values are mentioned below. Ultimately a comprehensive database of acceptable initial values would need to be compiled to address the issue of missing uncertainty information and metadata making estimation a necessity.

2.6.2.1 Worst Case Initial Elevation Uncertainty

2.6.2.1.1 Soundings Lead-line soundings can be assumed to suffer from vertical rounding error at all depths, this compounds the assumption that vertical uncertainty is in the order of 5% of depth. Horizontal accuracy depends on the technology used to determine position [10].

Sonar soundings are not subject to rounding uncertainty. Sounding data around Svalbard were assigned uncertainty in position of 0.1 - 10Km and in elevation of 5% of depth (one data set was assigned as 2% of depth) [80].

2.6.2.1.2 Photogrammetry accuracy can be difficult to determine. A method for determining the TPU of photogrammetry has been documented. In that instance uncertainty figures relating to the determination of the shoreline were assessed at about 3.3 meters. The system is currently location and configuration dependent [127]. This figure can provide a rough approximation of the expected uncertainty of the photogrammetry derived elevation data under similar circumstances.

2.6.2.1.3 Electronic and Digitized Charts tend to have a high level of uncertainty. The uncertainty propagation of the original data collection process and the uncertainty inherent in defining each isobath is compounded by reversing the isobath creation process to extract the information in point form (digitizing) [45]. The uncertainty assigned to position and elevation data extracted from isobaths is significantly impacted by the slope of the terrain where a data point is extracted from the chart. In steep terrain, between measured elevation points, an isobath location can be determined with greater precision. Point data extracted from isobaths created prior to the advent of modern positioning technology have been assigned uncertainty in position of up to 12 kilometers and in elevation of 5% of depth [80]. This elevation uncertainty can translate into 50 meters or more [34].

2.6.2.2 Initial Position Uncertainty

2.6.2.2.1 Long Range Navigation System (LORAN) was a radio navigation system that began in the 1940's and widespread use started in the 1950's. Loran A and C were both used extensively. Positioning using Loran C has an error that varies with time and geographic location [80]. The Loran C accuracy was in the order of hundreds of feet (in the region of 100 to 1,300 feet). The system remained active in North America until 1980, and the last system ceased service in 2000 [78].

2.6.2.2.2 Global Positioning Systems (GPS) steadily improved positioning accuracy. The positioning technology utilized has a significant impact on initial position accuracy. GPS has provided positioning information since 1964 and a variety of accuracy information is available to infer initial position uncertainty [89, 90, 106].

Accuracy information is available online for various other positioning systems (e.g. GLONASS [51], Galileo [128]).

2.7 Gridding Methods

Gridding methods are algorithms that use non-uniformly distributed data to create a uniform output. The iDEM system will use non-uniform elevation points to produce a DEM representing terrain.

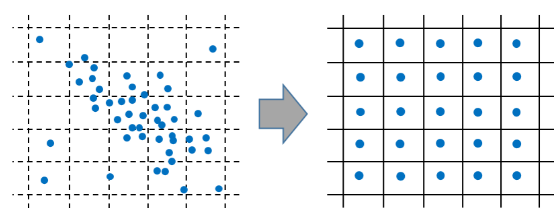


Figure 2.18: Point based gridding.

2.7.1 Quantitative Gridding Attributes

Inherent attributes can be used to compare gridding methods in a quantitative manner. These attributes are briefly defined below using the following terms [37]:

sample	-	a measured value
sample area	-	the spacial area (2D or 3D) in which samples are taken
predict	-	the calculation of a value at a particular spacial location based on existing sample values within a sample area
prediction surface	-	a surface defined by a collection of predicted values
zone	-	a smaller defined area contained within a sample area

In Figure 2.18 the left half of the diagram shows a sample area with sample points, and the right side of the diagram shows the predicted value locations.

Attributes define a gridding method to be one of the following [37]:

- Point or Area.
  - Point based methods predict point values at specific locations in space based on values of other points in space (see Figure 2.18) [37].
  - Area based methods predict a zone's value based on data from other zones (see Figure 2.19)[37].

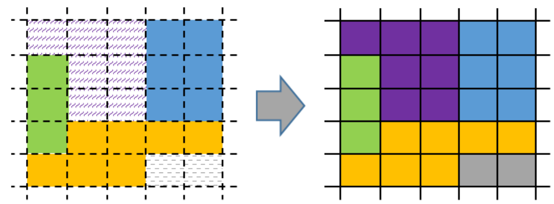


Figure 2.19: Area based gridding.

- Global or Local.
  - A method is *global* when it is used to predict the results for the entire sample area. If the value of one sample in the area is changed then the effect of reapplying a global function will be to alter all the predicted results (see Figure 2.20)[37]. In this example a search radius R is longer than the sample set. Altering the value of a sample point will effect all results.

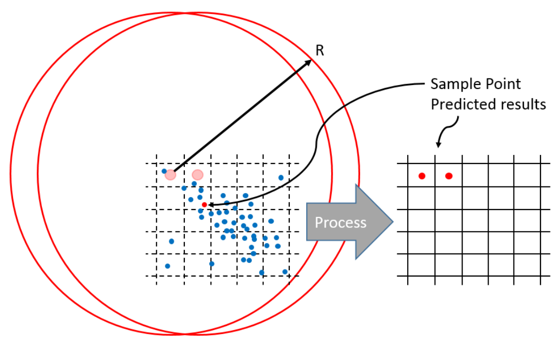


Figure 2.20: Global gridding.

– When a sample area is divided into subsets that are evaluated individually then the method is *local*. Changing a sample value and reapplying a local function will only affect the subset of predictions that use that value (see Figure 2.21)[37]. In this example local subsets are defined by radius r. Altering the value of a sample point will only effect results that use that sample point.

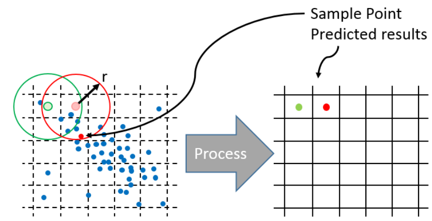


Figure 2.21: Local gridding.

- Exact or Approximate.

– If the prediction surface intersects the sample points then it is an *exact model* (see Figure 2.22)[37].

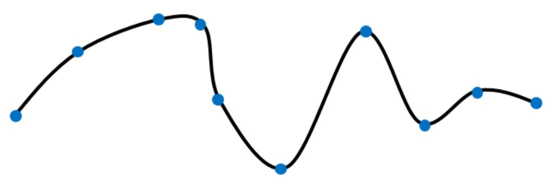


Figure 2.22: Vertical section example of an exact surface.

– In an *approximate*, or *smoothing*, model the surface passes near the sample points but does not necessarily pass through them (see Figure 2.23) [37].

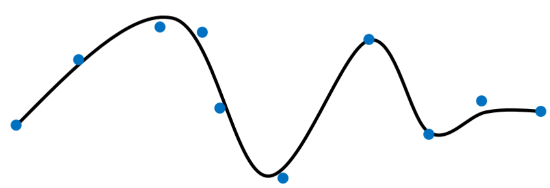


Figure 2.23: Vertical section example of an approximate surface.

- Stochastic or Deterministic.
  - Stochastic methods use the concept of randomness in the interpolation. This allows uncertainty of the resulting DEM values to be calculated. *Kriging* is an example of such a method (see 2.7.2.9). It assigns weights to the sample points based on the distance between points and the spatial autocorrelation among the sample points. The *autocorrelation* is determined by modelling the variability of the sample points as a function of the distance between them [37].
  - Deterministic methods use formulas or relationships to interpolate. Inverse Distance Weighted (IDW) interpolation assigns weights to the sample points based on distance (see 2.7.2.1). Unlike Kriging, IDW is deterministic because it doesn't incorporate autocorrelation randomness [37].
- Gradual or Abrupt.
  - Gradual interpolation creates a relatively smooth surface (see Figure 2.24)[37].

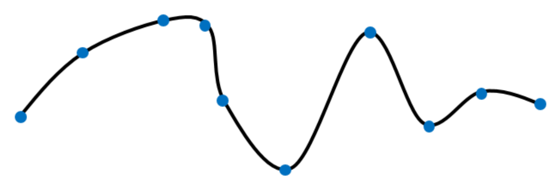


Figure 2.24: Vertical section example of a gradual gridding method.

  - Abrupt gridding will respond to changes by producing discontinuities (see Figure 2.25)[37].

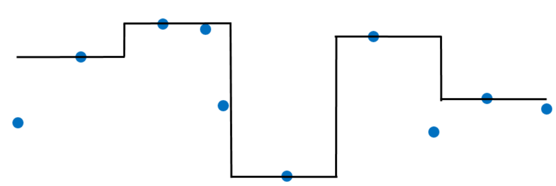


Figure 2.25: A vertical section example of an abrupt gridding method, the nearest neighbour method (see Section 2.7.2.3).

- Free or Constrained.
  - *Free* models can interpolate values beyond the range of the sample data (see Figure 2.26)[37].

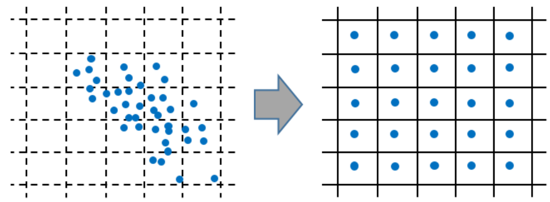


Figure 2.26: Example of a gridding method extending beyond the bounds of the sample data.

  - *Constrained* models are unable to predict values outside the range of the sample data (see Figure 2.27)[37].

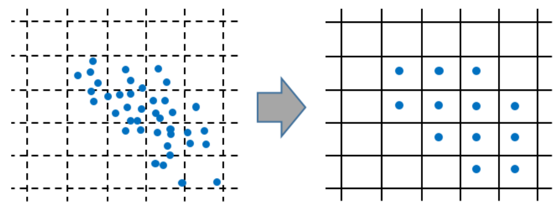


Figure 2.27: Example of gridding method constrained by the bounds of the sample data.

- Includes covariates.
  Additional variables can sometimes produce a better surface prediction. The additional information can help define the surface in greater detail or reduce the uncertainty. This is not usually applicable to DEM [37].

2.7.2 Comparing Some Common Gridding Methods

This section will review some commonly used interpolation methods [126]. Table 2.2 compares the selected gridding methods by summarizing the attributes that define each method. Note that the majority of the methods are deterministic; these tend to be less computationally intensive but uncertainty is difficult to determine [37].

	Global or Local	Exact or Approximate	Stochastic or Deterministic	Gradual or Abrupt	Free or Constrained	Includes Covariates
Gridding methods						
Inverse Distance Weighted	G/L	E	D	G	C	No
Natural Neighbour	L	E	D	G	C	No
Nearest Neighbour	L	E/A	D	A	C	No
Moving Average	L	A	D	G	C	No
Triangulation with Linear Interpolation	L	E	D	G	C	No
Minimum Curvature	L	E	D	G	F	No
Splines in Tension	L	E	D	G	F	Yes
Locally Weighted Smoothing Regression	L	A	S	G	F	No
Kriging	L	E/A	S	G	F	Yes
Combined Uncertainty and Bathymetric Estimator	L	A	S	G	C	No

Table 2.2: Attributes of some common gridding methods.

Brief descriptions of the selected methods are provided in the following sections. The knowlegable reader may wish to skip to Section 2.8.

2.7.2.1 Inverse Distance Weighted (IDW)

Also known as *Inverse Distance to a Power*, IDW is a weighted average interpolator. It can be exact or smoothing. The points closer to the prediction are allocated more weight in the averaging. Depending on how the algorithm is implemented the extent of the smoothing can be controlled. The relative influence of distant points can be altered by the choice of the power used in the weighting function. A higher power will reduce the influence of distant points while a lower power will increase it. Increased influence of distant points results in a smoother surface. Deterministic averaging methods are constrained so that the resulting interpolated values cannot exceed the range of the known data points (see Figure 2.28). The averaging nature of the formula flattens the predicted values beyond the sampled area to the mean value of the sampled area. The uncertainty contribution of the interpolation method is not easily obtained with this method [37].

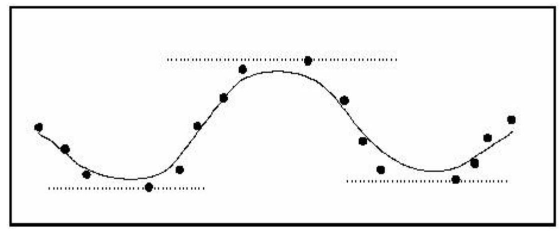


Figure 2.28: Inverse Distance Weighted Curve. [100]

2.7.2.2 Natural Neighbour Method

The Natural Neighbour Method is exact but the resulting surface is not smooth; in fact it is based on Voronoi tessellation and produces a series of polygon plates with distinct borders (see Figure 2.29). Polygons are defined such that the border between two points will be the midpoint. In a multi-point system the polygons are irregularly shaped. The uncertainty contribution of the interpolation method is not easily obtained with this method [37].

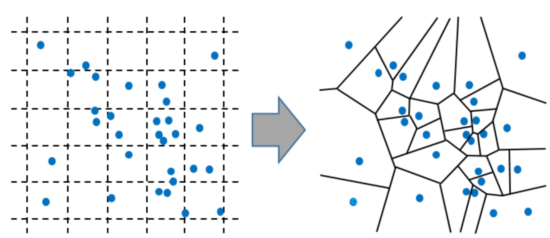


Figure 2.29: Natural Neighbour Method.

2.7.2.3 Nearest Neighbour Method

This is a fast method that assigns the value of the nearest data point to a grid node (see Figure 2.30). This method can be useful in previewing a surface or in filling in missing values in an almost complete grid. The uncertainty contribution of the interpolation method is not easily obtained with this method [37]. The value of each grid node is equal to the nearest sample value. The circles show the distance to the nearest sample point. Red points are unused.

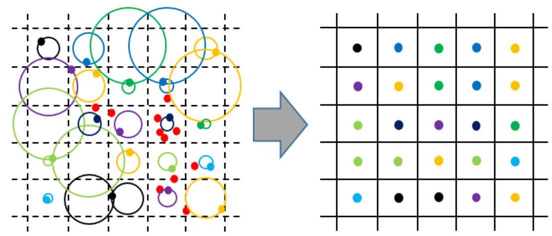


Figure 2.30: Nearest Neighbour Method.

2.7.2.4 Moving Average Method

This is not an exact method. A search radius is defined along with a minimum number of data points to average. If the required minimum number of points is found within the radius of the grid node then their values are averaged to compute the value at the node. If not enough points are found within the radius a blank value is assigned to the grid node (see Figure 2.31). The uncertainty contribution of the interpolation method is not easily obtained with this method [37].

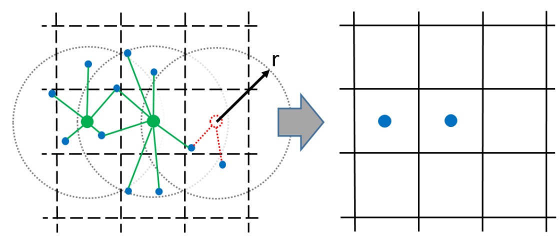


Figure 2.31: Moving Average Method with search radius (r) and support threshold of 4 points.

2.7.2.5 Triangulation With Linear Interpolation

This is an exact interpolator that uses triangular surfaces created between the data points. The surfaces are created using Delaunay triangulation where the points are joined by lines to create triangles so that no triangles overlap. This is an exact method, every data point is on the triangular surface. The grid elevation values are linearly interpolated on the triangle surfaces (see Figure 2.32). The uncertainty contribution of the interpolation method is not easily obtained with this method [37].

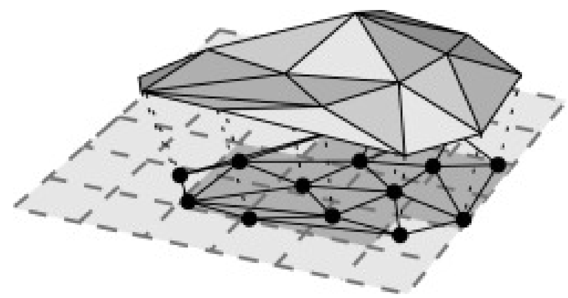


Figure 2.32: Triangular interpolation and DEM [70].

2.7.2.6 Spline Interpolation

Also known as *minimum curvature* this method is most commonly visualized as an elastic thin plate surface that is deformed to pass through all the data points. The deformation is performed so that the curvature of the surface is kept to a minimum (see Figure 2.33)[118]. In this Figure three parts of the curve are highlighted by curvature circles. Minimum curvature, spline interpolation, fits the smoothest possible surface by maximizing the circle sizes.

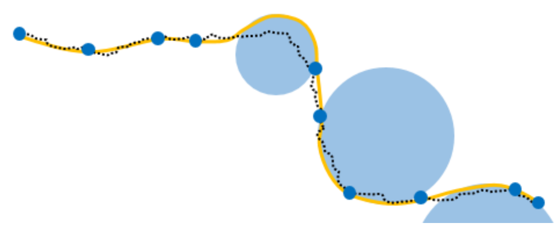


Figure 2.33: Spline interpolation. The dotted line is the actual terrain, The solid orange line is the fitted spline. The large circles of curvature illustrate minimal curvature.

There are two major drawbacks with this method:

- Uncertainty values are not inherent output of the gridding method [118].
- The minimum curvature method does not approximate topography or bathymetry well [118].

2.7.2.7 Splines in Tension

The difficulty of minimum curvature Spline Interpolation to model topography and bathymetry well can be improved by adding a tension term to the function. The tension term relaxes the requirement for minimum curvature, allowing closer correlation of the resulting surface and the real terrain (see Figure 2.34). The tension term is a variable with a range of 0 to 1. Experienced terrain modellers have determined that the tension term should generally be set at a value between 0.25 and 0.35 depending on the condition of the terrain being modelled [86, 118]. Splines-in-tension is a relatively easy method to use making this a common way of creating DEM. The method is implemented in publicly available software such as Generic Mapping Tools (GMT) [125] and MB-System's MB-ZGRID function [16]. In this Figure three parts of the curve are highlighted by curvature circles. The tension term in this spline algorithm allows the curvature to be customized. The curvature can be substantially increased as indicated by the small circle sizes.

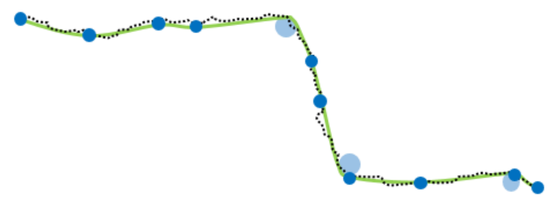


Figure 2.34: Splines in tension. The dotted line is the actual terrain, The solid green line is the fitted tension spline. The small circles of curvature illustrate higher than minimal curvature.

Spline interpolation has some difficulty with sparse data and with large variations in data density. A method utilizing the stacking of splines-in-tension created at different resolutions employs a masking technique to merge the results. This method deals well with data density variation improving the result over

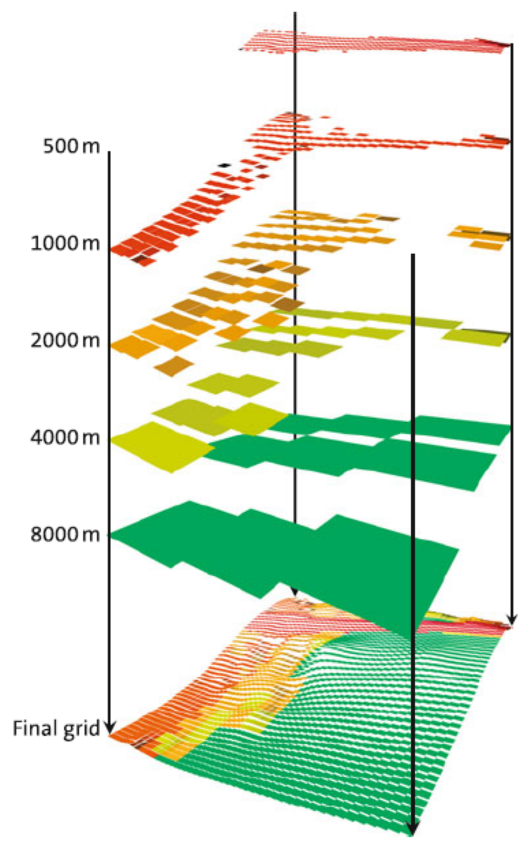


Figure 2.35: Stacked spline-in-tension [57].

the global solution of ordinary splines-in-tension (see Figure 2.35) [57]. This method is implemented in GMT.

The optimal choice of the tension can be difficult to determine. The best value for the tension term cannot be calculated based on the data. Methods do exist to optimize the tension term by repeated gridding using different tension values to determine the best fit [86]. For the best result, it may be that the area of interest needs to be subdivided into smaller DEM so that different tension values can be utilized for different terrain types.

Figure 2.35 illustrates the stacking of spline-in-tension grids with different resolutions. The resolution of each grid to be stacked is determined by the data density of the area modelled.

The uncertainty contribution of the interpolation method is not easily obtained with this method [37].

2.7.2.8 Local Weighted Smoothing Regression (LOESS)

is also known as Weighted Quadratic Least Squares Regression and as Locally Weighted Polynomial Regression (see Figure 2.36)[21, 22]. This method combines the simple Least Squares Regression with the flexibility of Non-linear Regression. There is no need to specify a global function with which to model the data [31, 88]. LOESS locally fits subsets of the data and uses them to build a function that describes the variation of data [23]. This method is computationally intensive [88]. LOESS is not exact, it produces a smoothed result that is useful in defining large scale trends or patterns [10, 40]. It is not an interpolator method as it does not predict unknown values. This gridding method is prone to aliasing where the data is sparse. The uncertainty contribution of this gridding method can be obtained in a simple manner [37, 88].

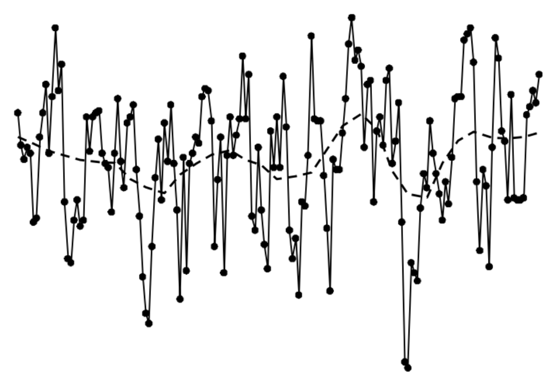


Figure 2.36: Example showing a solid line that exactly fits the data and a dashed line depicting a LOESS smoothing of the data [31].

2.7.2.9 Kriging

This is a stochastic geostatistical method that can be exact or smoothing. In general it computes the weighted average of known values of a function in a general area to predict the value of the function at a specific point. This is the same general idea as the IDW method. The IDW method only uses the distance between the points to weight the influence of the surrounding data.

A variogram is a function describing a spatial random field in terms of dependence. For DEM it is defined as the variance of the difference between the values at different locations [7]. In terrain modelling the elevation values at two sample locations will vary with the distance between locations. Elevation measurements taken far apart will vary more than measurements taken close together. Different variogram models can be utilized. In terrain modelling the common variogram models tend to be spherical, exponential or Gaussian. Variogram models have different ranges of influence; a model's range of influence the the area over which the model can be utilized accurately. There are indications that in terrain modelling the spherical variogram model has a larger range of influence, up to 3°[6]. That is the spherical model variogram may be accurate as far as a geographic distance of 3°. The largest DEM that can be created using a specific variogram model is limited to its range of influence. Isotropy implies directional independence, that is identical properties in all directions. Anisotropy implies the opposite, direction dependence. The terrain in a given area of the globe often exhibits anisotropy, such as ocean floor ridges running parallel in a given direction. The variogram and anisotropy effectively customize the resulting surface using the statistical factors describing the terrain [84].

Unlike IDW, Kriging incorporates anisotropy and uses a variogram to create a custom fitted surface. The statistical nature of this method allows it to generate an uncertainty map of the resulting surface [86].

Three different levels of sophistication are possible with the Kriging method. Simple Kriging assumes a known stationary mean which is generally not true for terrain modelling; the mean of a location changes as distance between locations increases. Ordinary Kriging extends simple Kriging by assuming the mean is unknown but that it can be to be estimated from the data. Universal Kriging extends ordinary Kriging by assuming that there is an underlying trend in the data that can be expressed using a polynomial model. For the creation of a DEM ordinary Kriging is usually the most appropriate option [84, 3].

2.7.2.10 Combined Uncertainty and Bathymetric Estimator

The Combined Uncertainty and Bathymetric Estimator (CUBE) is primarily designed for high density data, like that obtained from MBES. The system attempts to isolate random errors. It does this by creating multiple potential surface solutions. The best solution can be determined by several methods but at present the fastest and most reliable method is to select the surface that contains the majority of points. In order for CUBE to function correctly systematic errors need to be corrected prior to processing. While CUBE will highlight existing systematic errors it will not compensate for them. [12]

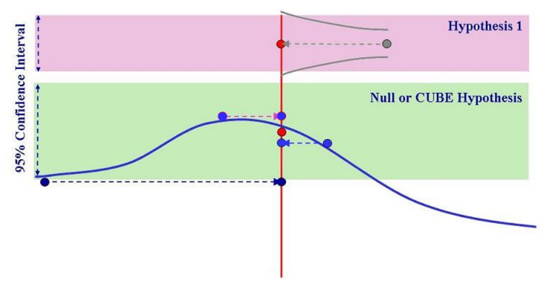


Figure 2.37: Soundings (blue) are projected to a grid node. Uncertainty is also projected and increased quadratically with distance. Projected soundings are sorted by overlapping uncertainty envelopes and each group forms a separate hypothesis. Two potential surfaces are displayed in this figure. The upper red point is the calculated depth at the grid node based on hypothesis 1. The lower red point is the calculated depth at the grid node based on hypothesis 1 [60].

The algorithm creates a regular grid of the desired size by searching for all soundings within the radius of 5% of the depth (see Figure 2.38). Those soundings are projected to the grid node as shown in Figure 2.37. The uncertainty is also propagated but it increases in a quadratic manner with distance. As with the IDW method soundings are effectively weighted by distance but CUBE does not simply use the distance of the sounding from the grid point. Uncertainty propagation means that soundings closer to a grid point will have lower uncertainty. CUBE uses this fact by assigning a weight that is inversely proportional to uncertainty, the lower the uncertainty the larger the weight assigned [60].

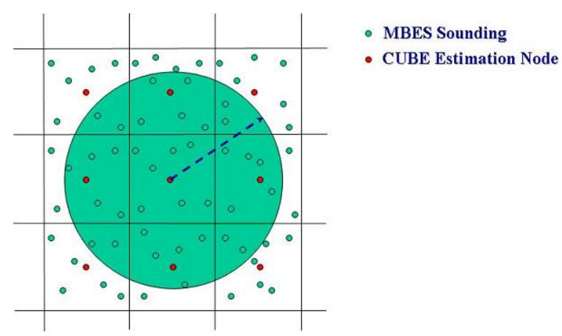


Figure 2.38: Search radius equal to 5% of depth. MBES soundings are the high density multi-beam sonar data points. CUBE estimation nodes are the grid nodes of the resulting DEM [60].

This algorithm recognizes that data may have been recorded from more than one surface, for example the ocean floor and a school of fish. It performs this task by comparing each projected sounding with all projected soundings at the grid node. If a projected sounding is within the uncertainty envelope of another projected sounding then the two soundings are assigned to the same surface hypothesis (see Figure 2.39). If a projected sounding is not statistically compatible with the existing depth estimate (not within an existing uncertainty envelope) it is separated into a new potential surface. This process means that judgement of which is the best surface is not made until all the soundings within the radius have been considered. [12]

Traditionally cleaning of data involves the removal of systematic and random errors that generate artificial artefacts in a DEM. This requires the manual intervention of an expert in the field to be truly effective. CUBE does not eliminate the manual intervention but it does provide a tool to greatly reduce the required work. [12]

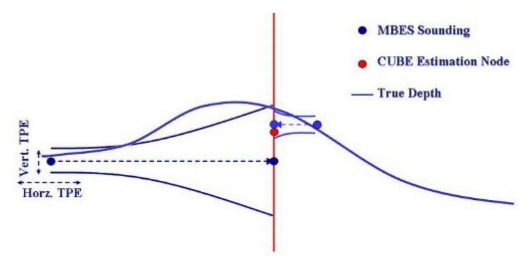


Figure 2.39: Depth and uncertainty projection. The two blue projected soundings are within the same uncertainty envelope so they are considered to be part of the same potential surface. Based on these two soundings the elevation value at the grid node is calculated to be the red CUBE Estimation Node [60].

CUBE is not an interpolation system so it does not calculate values for any grid nodes that do not have enough direct evidence (soundings) to support them; it leaves such nodes blank. CUBE does provide a grid of available data at the desired resolution with uncertainty. The uncertainty is calculated using a method that is not computationally intensive [12].

2.8 Some Existing Fusion Methods

Stitching existing DEM together to form larger models can be considered a simple fusion technique. The edges of DEM are usually less accurate than the interior because there is usually no data available beyond the edges to accurately define them. This usually makes the joins between the DEM areas of distortion where discontinuities and artifacts form. The joins, or seams, between perfect DEM that have dissimilar resolutions generally suffer from resolution distortion, another cause of discontinuities and artifacts. Feathering of the seams is a method of smoothing discontinuities along joins. An example of a feathering technique is averaging the joined DEM points along the join by weighting the influence of a point depending on the distance from the join

[114]. These sorts of solutions were the first attempted to correct these issues, and the distorted edge terrains of the two DEM were blended at the join. This usually creates ramps or slopes at the joins, effectively smoothing the area, but this is rarely an accurate depiction of the terrain [114, 54].

A DEM can be considered to be a preprocessed dataset of elevation points. Rather than feathering, a better solution is to fuse the DEM as though they were datasets of elevation points and calculate new DEM from the merged data. This would eliminate discontinuities and reduce artifacts. This method will not be as effective as fusing the original cleaned datasets that were used to create the DEM. The data points extracted from the edges of the DEM to be joined inherit the edge distortions created when the DEM were calculated.

At this point in terrain modelling the uncertainty of the resulting surface is becoming an attribute of greater interest. Most of the older and some of the newer data does not have associated uncertainty information but methods exist to estimate such values [79, 39, 38]. The uncertainty information can be utilized to improve fusion techniques and to highlight the areas of DEM where artifacts are most likely to occur.

2.8.1 DEM Tiling

A common source of artifacts is the seam where DEM covering adjacent areas are joined. One way to eliminate these joins is to create one large DEM covering an extensive area. An issue with this approach is that the resulting file is large.

Another issue is that many uses of DEM only require a limited DEM area to be employed during any one particular operation. One way to address this issue is by extracting the desired data from the larger DEM.

The most significant issue with the use of large DEM files is updating those files. New data from accurate soundings such as MBES surveys generally do not cover an extensive area. This means that an update to an extensive DEM only changes a small portion of the DEM but requires the whole DEM to be recalculated to integrate the new data.

One way to avoid the extensive recalculation issue is to tile the data. If an extensive DEM is chopped up into a regular grid of tiles then only the tiles affected by the new data need to be updated. The issue with updating individual tiles is that such updates usually distort the surface at the edges of the tile. This distortion is not caused by the new data but rather results from a lack of data beyond the edge of the tile. This lack of data limits the accuracy of the calculations of the tile edges of the recalculated DEM. This issue leads back to the initial issue in that the seams between the tiles now become sources of artifacts.

An alternate approach to storing large DEM is named Global Multi-Resolution Topography synthesis. This method implements the overlapping of tiles which effectively blends datasets. At a given resolution the large DEM is stored as overlapping tiles [112]. This overlapping tile method allows a DEM region to be updated without updating the whole model. The overlapping ensures that gradient calculations can be accurately performed on data at the edges of the tile eliminating distortion due to lack of data. Figure 2.40 shows the corner of a tile with a boundary that overlaps neighbouring tiles.

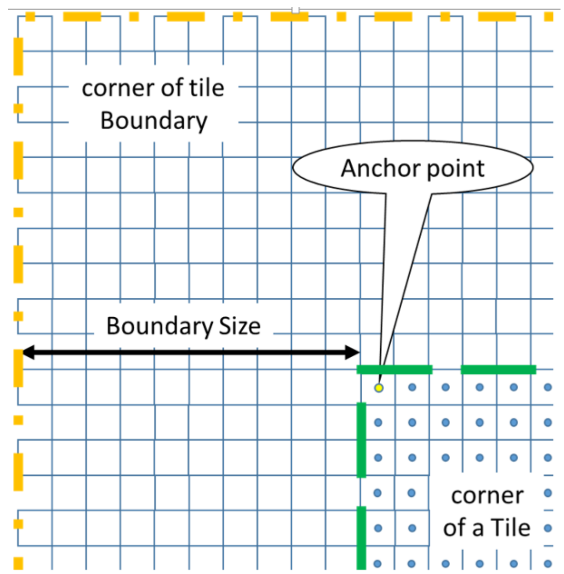


Figure 2.40: Tile boundary.

2.8.2 CUBE for High Density Data

Several applications exist to process the rapidly growing MBES data, such as MB-System, CARIS, Fledermaus [93]. There are a variety of different levels of processing possible with these systems. Some provide automated outlier detection but most also require extensive expert operator interaction. There is an ongoing effort to continue the development of high-end semi-automated processing systems.

At present NOAA has accepted the Combined Uncertainty and Bathymertry Estimator (CUBE). With this acceptance most of the major existing applications are incorporating CUBE into their suite of processing capabilities. CUBE's ability at semi-automated outlier removal that provides grids with uncertainty makes it the primary candidate to handle high density data. Assuming that systematic errors have been eliminated CUBE will blend high density datasets in such a way that the fused resulting DEM is minimally affected by artifacts [18, 19].

It should be noted that the resulting DEM will not be populated where a lack of high density data does not support the CUBE method. Other algorithms need to be applied to process lower density data regions where interpolation is necessary to fill in empty DEM nodes.

2.8.3 Three Step Approach for Lower Density Data

In 2008 Paul A. Elmore and Chad A. Steed recommended a three step approach to the fusion of gridded and sounding archival datasets (see Figure reffig:TSA2008) [40]. This is based on Brian Calder's solution to creating DEM from archival soundings with associated uncertainty (see Figure reffig:TSA2006) [10]. This method will be referred to as the Three Step Approach or 3Step for short in this text.

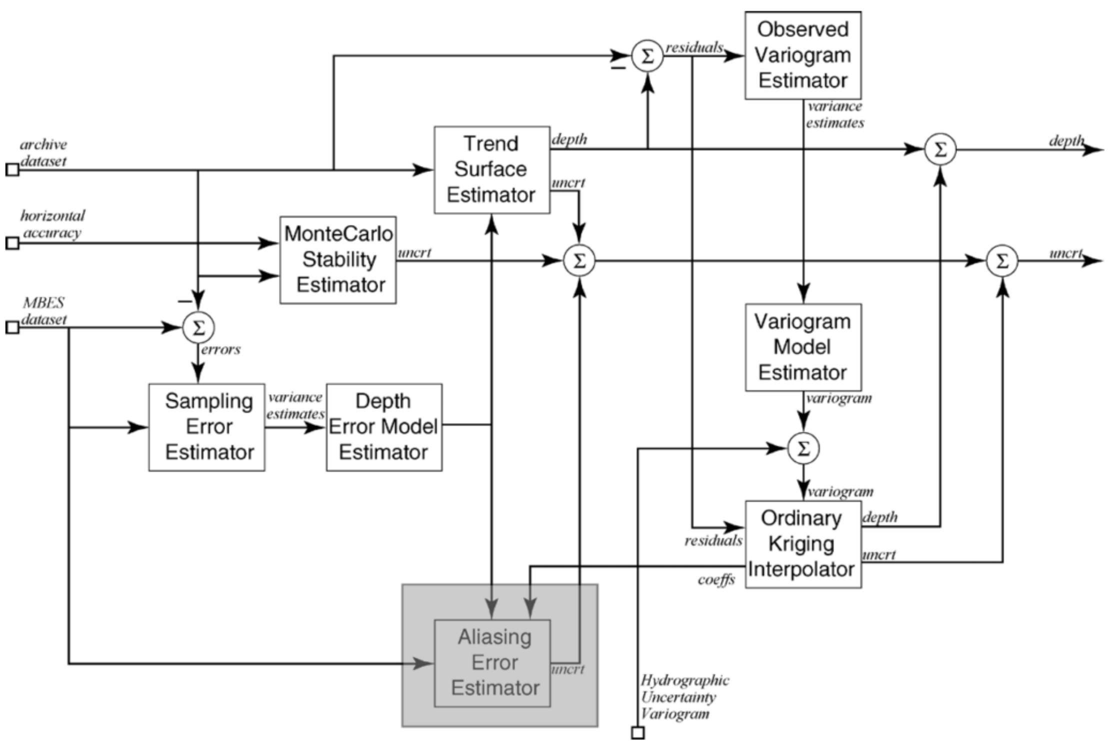


Figure 2.41: Calder's 2006 Three Step Approach to fusion of archival soundings [10].

The 3Step ingests archived datasets to fuse them into a gridded DEM with associated uncertainty. The archival data points can be soundings or gridded DEM. The sounding data requires associated horizontal uncertainty values and the gridded data is assumed to have associated vertical uncertainty. The resulting DEM are produced from sparse data. This method can, but is not designed to, process high density data produced by MBES, which usually have associated uncertainty values. Horizontal uncertainty estimates can be input to the 3Step to estimate the archival data measurement error.

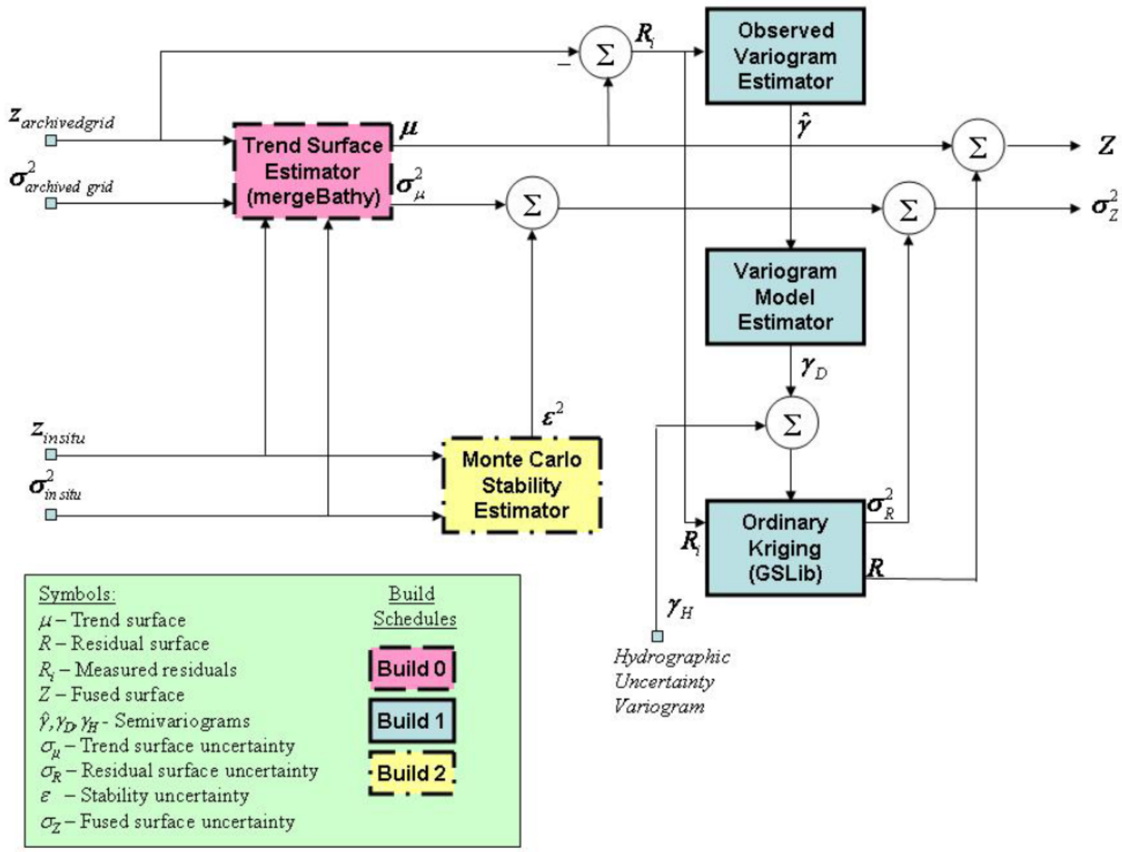


Figure 2.42: Elmore and Steed's 2008 Three Step Approach to fusion of archival grids [40].

2.8.3.1 Step 1: Create a Trend Surface.

A trend surface can be created at the desired grid size from the archival grid or sparse data. This is done to take into account the spatial variation in the data and to provide a surface for comparison with available high quality estimates to ascertain archival measurement errors. Both the 2006 (see Figure 2.41) and 2008 (see Figure 2.42) versions of the 3Step use LOESS quadratic interpolation with a recommended search radius (d0) set to ten times the average sample distance of the data.

2.8.3.2 Step 2: Restore the Surface Detail.

The LOESS interpolation produces a smooth trend surface lacking the finer detail embedded in the original data. This detail can be restored by determining the difference between the original data points and the trend surface. The resulting new surface of residuals can then be interpolated using Ordinary Kriging. The trend surface and the residual surface are added together to produce a detailed grid surface. With the assumption that statistical independence applies, the variance of the trend surface and that of the residual surface can simply be added together [40].

2.8.3.3 Step 3: Estimate Additional Uncertainty.

Additional uncertainty related to the horizontal positioning error needs to be estimated. In both versions of 3Step a Monte Carlo uncertainty estimator is employed using the trend surface and residual Kriging method of gridding [79]. The resulting variance is also added to the uncertainty surface to produce a total uncertainty. This total uncertainty does not take into account potential aliasing contributions.

As part of the iDEM design, which the next chapter discusses, the three fusion methods mentioned here (Tiling, CUBE, and 3Step) have been amalgamated into a new, broader, solution than each provides individually. The tiling method of storing data is embedded into the data storage and retrieval system to reduce processing distortion and enhance processing efficiency. The data fusion stage of iDEM utilizes an amalgamation of CUBE and 3Step. Other methods that are efficient for Gaussian data fusion can be incorporated as additional options to be used instead of the existing amalgamation. The iDEM system is designed to allow each distinct function to be replace with other algorithms that utilize the same input and produce the similar output. This research utilizes the CUBE and 3Step amalgamation as an example of a possible fusion method.

Chapter 3

The New Approach and Solution

Many applications already exist that manipulate, analyse and process elevation data. The majority of these applications are mature products with abundant features. They do not have a specific purpose but are feature rich with the ability to transform elevation data into a variety of different products from statistical analysis results, to contour maps, to DEM. The applications usually contain substantial functionality designed to *clean* raw input data by assisting with the identification and removal of random and systematic error. However the expert user can interpret the results of the application features to identify and eliminate errors. Such applications usually have multiple algorithms available to create a DEM. The versatility of the abundant features require a trained user to utilize them to their potential, to navigate the applications efficiently, to understand the information displayed by the applications and to decide how to produce optimal output. These applications are, inherently, labour intensive.

A Combined Design

The iDEM system has a single purpose and is designed to be feature lean with automation to minimize required user expertise and interaction. Data cleaning will not be a central concern for this version, and the product quality will be largely dependent on the input data quality.

The iDEM system can import cleaned data, existing DEM and raw elevation data as input, and uses an amalgamation of the CUBE and Three Step Approach methods of gridding to create a DEM with DCM output. The system accomplishes the process of transforming input data into DEM and DCM in three distinct stages:

- Homogenize Input Data - translates diverse points into homogeneous datasets

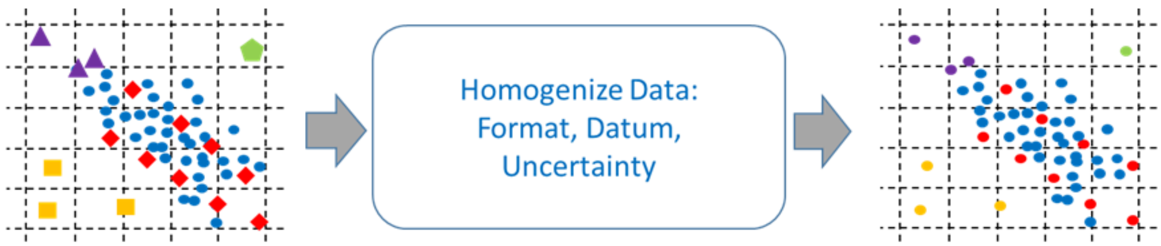


Figure 3.1: iDEM Stage 1 concept.

- Merge Datasets - merges homogeneous datasets into a single set

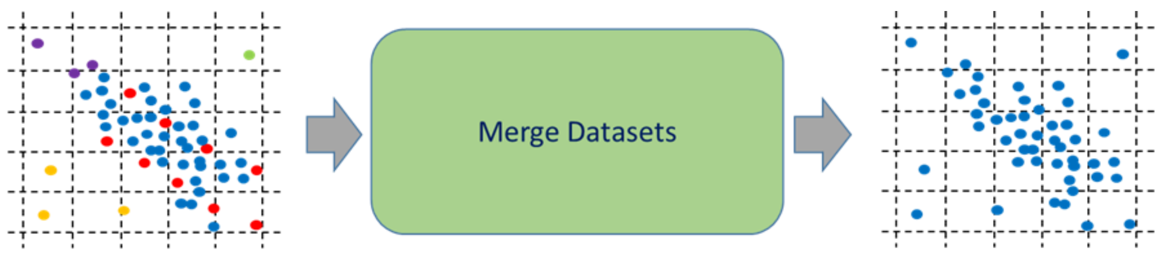


Figure 3.2: iDEM Stage 2 concept.

- Grid and Interpolate - transforms single set data into DEM with DCM

Figure 3.3: iDEM Stage 3 concept.

3.1 Stage 1 - Data Homogenisation

Stage 1 of iDEM imports external bathymetric and topographic datasets in most common file formats. The input data types are listed in order of preference below.

Datasets are comprised of:

- cleaned elevation data with uncertainty,
- DEM with uncertainty,
- cleaned elevation data without uncertainty,
- DEM without uncertainty,
- raw elevation datasets.

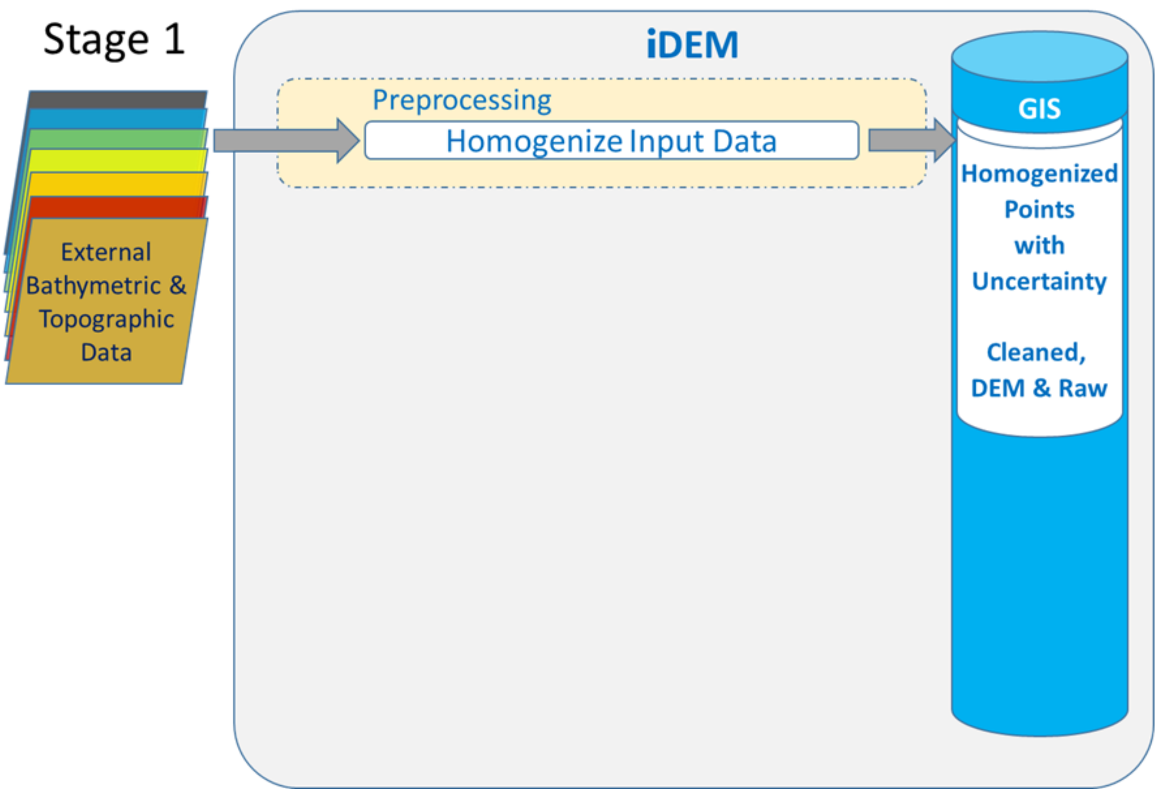


Figure 3.4: iDEM Stage 1 detail.

The data is homogenized to ensure that all the data stored in iDEM's relational database have the same coordinate system, position datum and elevation datum (but not tidal datum). Any datasets that are missing uncertainty information will be processed to add estimated uncertainty for each data point within the dataset.

In Figure 3.4 the GIS crown on the relational database cylinder indicates that the database has a spatial Geographic Information System extension. That extension increases the database functionality by adding modules to store, retrieve, analyse and manipulate spatial data. Some of this functionality is utilized in the homogenization of the input data. The ability to handle spatial data using SQL commands within the database simplifies the design of iDEM.

3.1.1 Coordinate System and Datum

Using the World Geodic System 1984 (WGS-84) allows data of any portion of the globe to be stored in the same format. WGS-84 incorporates a global coordinate system, horizontal datum and vertical datum. Together these components of the standard allow any position on the globe to be defined as a tuple: (longitude, latitude, elevation).

3.1.2 Tidal Datum

Various uses of DEM require different reference for the sea level datum. Nominal sea level is not always the best option for creating a DEM with a particular use in mind. For example tsunami modelling often uses Mean High Water (MHW) level or (in regions of highly variable tides) even Mean High High Water level (MHHW). These varying tidal datum requirements mean that the DEM sea level datum should be adjustable after the DEM has been created. The iDEM system can store a nominal sea level DEM (DEM_{nsl}) and an adjustment algorithm that includes that DEM section of the globe. When a DEM with a certain datum is desired a copy of the DEM_{nsl} can be adjusted to the required datum, e.g. an MHW DEM. This allows optimization of the DEM sea level datum for the intended task without undertaking the computationally intensive task of recreating the DEM from input data.

Ideally the homogenized data should store elevation measurements referenced to the nominal sea level. This would allow tidal adjustments to be applied for any region of the globe in a single step. Under such an ideal system the assumption that tidal models are available for the whole globe comes into play. Such models would be required in order to homogenize the data. A limitation of such a default sea level standard for iDEM would be the prevention of loading any data for any region that lacks a tidal model unless the data is actually referenced to nominal sea level (EGM96). Many existing datasets have sea level datum optimized to the task they were created for.

It is possible to create a DEM without a common sea level reference. The accuracy of such a DEM would be compromised by such a compilation but sometimes inaccurate data is better than no data, especially if a DCM accompanies the DEM to highlight the fact.

To minimize restrictions iDEM input data does not require transformation to a common sea level datum. The input datum will be recorded so that at the time a DEM is compiled of a particular area the datasets being utilized can be adjusted to a common sea level datum at the time of compilation. Such adjustment is contingent on a tidal model being available for the area. The user will specify if iDEM should proceed without tidal adjustment.

3.1.2.1 Tidal Inference

Some datasets are not referenced to a tidal datum. For bathymetry in deep waters this is generally not an issue as the uncertainty of such elevation measurements is greater than the tidal variation. In such cases the tidal datum can be assumed to be nominal sea level.

When a dataset in shallower waters lacks a tidal datum and that dataset overlaps or adjoins another dataset with a tidal datum it is possible to infer the missing one. The greater the number of overlapping or adjoining data points with tidal datum the greater the accuracy of the inference. This process is only worthwhile if the dataset with unknown tidal datum has an uncertainty smaller than the local tidal variations. If the uncertainty is greater than the local tidal variations, nominal sea level datum should be assumed.

This inference process may also be applicable to estimating a tidal model of an area. If an area overlays or adjoins one or more known tidal model(s) a local model can be inferred. The accuracy of the inferred model will be compromised by the fact that tidal variations are not just time dependent but are also affected by local terrain, currents, etc.

Tidal inference is not strictly required to create DEM but it is a part of the homogenization of the data. It can be considered as a data cleaning process that reduces a systematic error.

3.1.3 Uncertainty Estimation

Elevation data that is used as input will not necessarily be accompanied by uncertainty information. In order to build a DCM for each DEM, uncertainty information is required. When such information in unavailable iDEM will estimate uncertainty values. iDEM will initially employ the simpler Monte Carlo approach (see Section 2.6.1).

A drawback of front loading a database with Monte Carlo uncertainty estimation is that the resulting uncertainty may not be the same at different grid sizes. The Monte Carlo process interpolates the data at the desired grid size of the resulting DEM [79]. It then uses that interpolated data to estimate uncertainty for the data point. Whether there are any significant differences in the estimated uncertainty of any single data point for varying grid sizes requires additional research. The iDEM system assumes that there are material differences in such uncertainty values at different grid resolutions so the system will save multiple uncertainty values for each data point. Data points that require uncertainty estimation will have one uncertainty value for each grid resolution. Saving these estimated uncertainty values decreases the computations required to grid and interpolate a DEM at the cost of requiring additional storage space.

3.2 Stage 2 - Merging Datasets

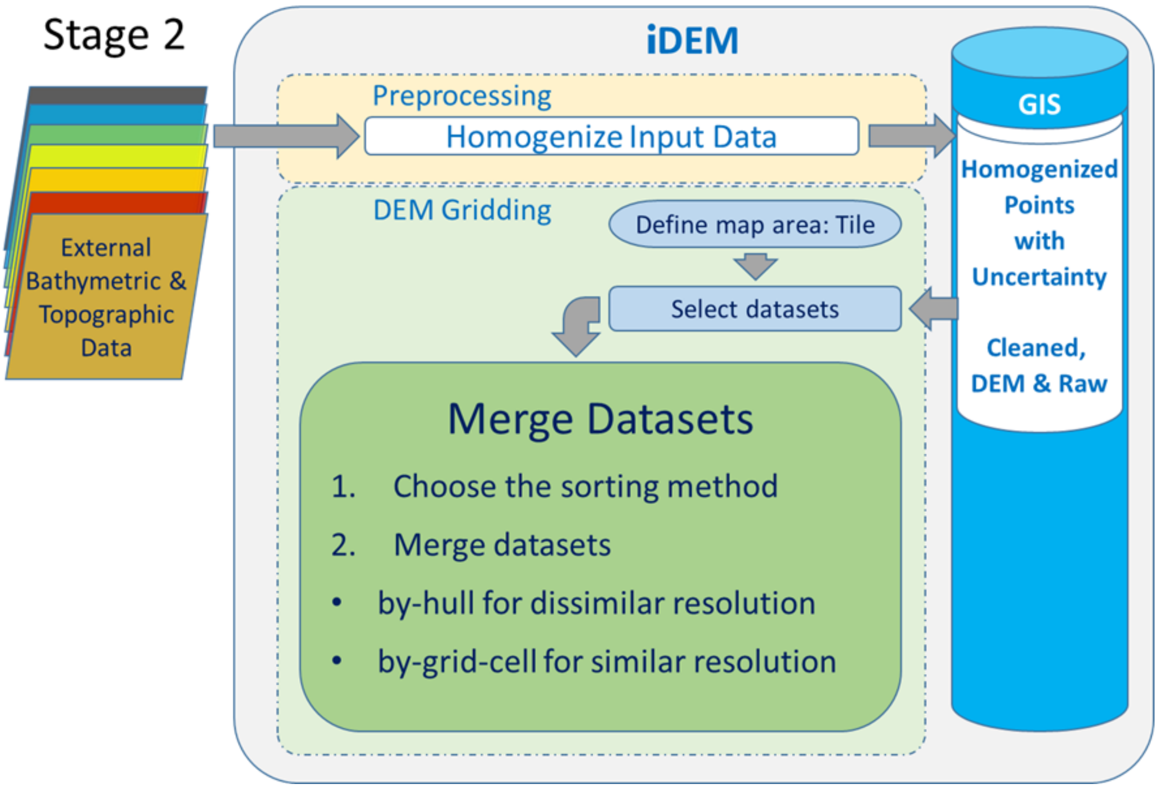


Figure 3.5: iDEM Stage 2 detail.

The Stage 2 overview shown in Figure 3.5, which depicts three steps in the merging stage:

Step a) Define map area: Tile

Step b) Select datasets

Step c) Merge Datasets

These three steps are described below.

3.2.1 Stage 2A Defining the Map Area: Tile

3.2.1.1 Grid Type

Grids can either be corner referenced or center referenced. In a corner referenced grid a value is assigned to node points where the grid lines intersect. In the center referenced grid a value is assigned to a node in the center of the grid cell bounded by the grid lines. The raster data format favours the center referenced grid system. To increase compatibility between vector points as stored by iDEM and raster grids, iDEM utilizes the center referenced grid system. A raster grid cell can simply be represented as a point in the center of the cell with an elevation value equal to the raster value.

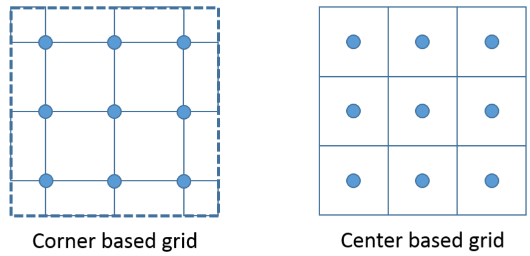


Figure 3.6: Node versus center grids.

Using the center referenced grid method with a common frame of reference for the points within the grid cell affects the relative positioning of tiles at the different resolutions. Two tiles covering the same area as represented by a DEM will be offset in position.

A consequence of this is that tiles of different resolutions will not tesselate. Figure 3.7 shows an example of two tiles at higher resolution with two lower resolution tiles superimposed. This figure shows the offset of tile positions even though there is clearly point co-alignment or position interlock (all lower resolution points co-align with higher resolution points). Using standard grid sizing (see Section 3.2.1.2.2 below) the maximum (x : y) offset is approximately (1,432 : 2,223) meters at 50° latitude and that only occurs by comparing resolutions of 144 and 1/9 arc seconds. Comparing 144 and 36 arc second grids reduces the offset to approximately (1,075 : 1,668) meters at 50° latitude. The offset is equal to half the difference between the two resolutions, making the offset smaller than half a grid cell. For a 1° × 1° tile the maximum offset is less than 2%.

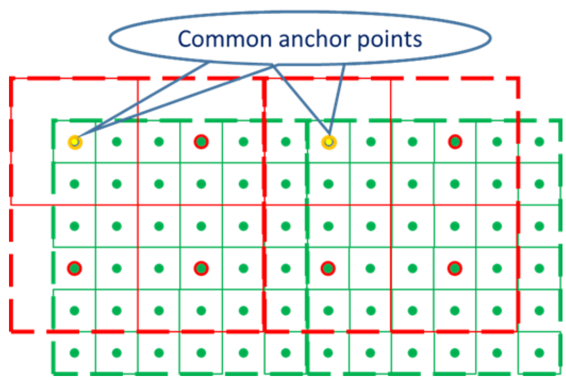


Figure 3.7: Tile resolution offset with point co-alignment.

3.2.1.2 Tile Definition

A DEM is actually an elevation map of an area. The iDEM system can store DEM of different sizes and at different resolutions. The overall vision is to store DEM covering wide areas in overlapping tiles of default sizes. It is possible to create custom tiles at will. In the discussion that follows a tile refers to both a portion of a larger DEM (that can be realized through tessellation) and custom DEM that may not have other tiles to tessellate. The area that a DEM is to cover will simply be referred to as a tile. In iDEM a tile is defined by giving values to several variables:

- anchor position defined as the [West, North] coordinates (latitude and longitude) from the top left corner in decimal degrees.
- tile size defined as the x and y size in arc seconds (default = 3600" = 1°)
- grid size in arc seconds
- list of datasets to utilize (default = [ALL], option to list dataset IDs)
- sorting method (default = ($Density \div Uncertainty$) see 3.2.3.2)
- similarity threshold (default = 3, utilized by merging algorithm see 3.2.3.5)

3.2.1.2.1 Optimizing the Tile Definition. There is a distinct advantage to having layers of varying resolution aligned (see section 3.2.1.1) so that the point positions interlock even though the tiles of differing resolution DEM are offset. The grid point positions of each lower resolution grid will precisely correspond to a grid point position on every grid of higher resolution (see Figure 3.9).

Therefore iDEM needs to ensure that the tile definition restricts the tile position and size so that different DEM resolution layers have anchor points that precisely overlap and so that only whole grid cells fit within the tile.

This means that the anchor position and tile size are not necessarily accepted by iDEM as entered. They are automatically adjusted to ensure the optimum conditions are met (see Figure 3.8). Both anchor position and tile size are linked to the intended grid size of the DEM. The anchor point must always be a whole multiple of the grid size from the nearest whole degree of longitude and latitude. In theory this means that the maximum adjustment is half a grid size in both x and y positioning for the anchor point and one grid size for the overall size of the DEM in both x and y dimensions. In practice an additional restriction is imposed by the standardizing of grid sizes as discussed below.

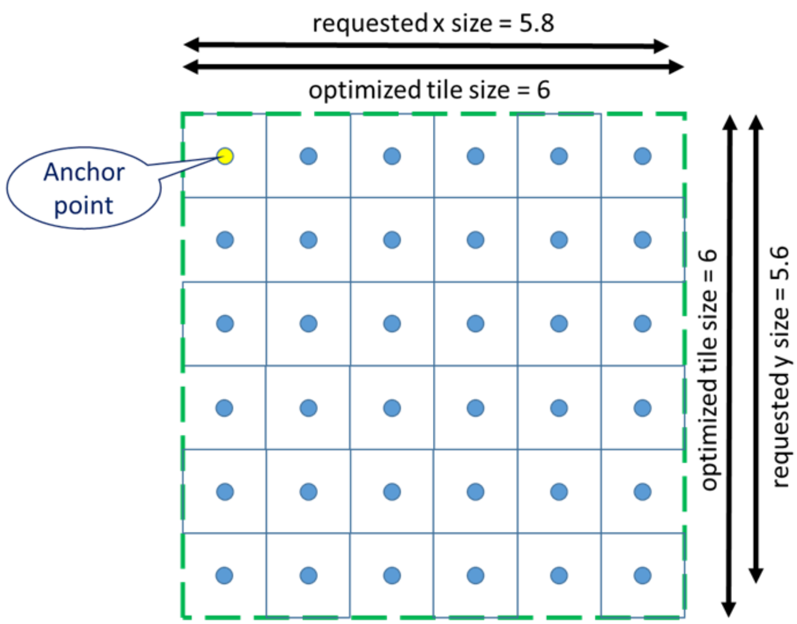


Figure 3.8: Optimized tile size.

Given:

lon_a	: *requested anchor longitude*	lat_a	: *requested anchor latitude*
lon_f	: *final anchor longitude*	lat_f	: *final anchor latitude*
gs	: *grid size (in decimal degrees)*		
$\lfloor A \rfloor$	$\equiv$ *round down to nearest integer*	$\lfloor A \rceil$	$\equiv$ *round off to nearest integer*
$A - B\lfloor \frac{A}{B} \rfloor$	$\equiv$ *A modulus B*	$(if\ B = 1) \Rightarrow$	$A - B\lfloor \frac{A}{B} \rfloor \equiv A - \lfloor A \rfloor$

$$lon_f = \lfloor lon_a \rfloor + \left(\left\lfloor \frac{lon_a - \lfloor lon_a \rfloor}{gs} \right\rceil \times gs \right) \tag{3.1}$$

$$lat_f = \lfloor lat_a \rfloor + \left(\left\lfloor \frac{lat_a - \lfloor lat_a \rfloor}{gs} \right\rceil \times gs \right) \tag{3.2}$$

The tile size parameters must always be a whole multiple of the grid size:

Given:

gs	$:grid\ size\ (in\ decimal\ degrees)$	$\lfloor A \rceil$	$\equiv\ round\ off\ to\ nearest\ integer$
x	$:requested\ tile\ size\ x\ direction$	y	$:requested\ tile\ size\ y\ direction$
x_f	$:final\ tile\ size\ x\ direction$	y_f	$:final\ tile\ size\ y\ direction$

$$x_f = \left\lfloor \frac{x}{gs} \right\rceil \times gs \quad (3.3)$$

$$y_f = \left\lfloor \frac{y}{gs} \right\rceil \times gs \quad (3.4)$$

Notice that when iDEM standardized grid sizes are employed the smallest gridSize used to adjust the tile size parameters should be 9 arc seconds (see Standardized Grid Sizes below).

A polygon defining the tile edges can be derived from these equations:

Given:

W	$:longitude\ of\ west\ edge\ of\ grid\ cell\ polygon$
N	$:latitude\ of\ north\ edge\ of\ grid\ cell\ polygon$
E	$:longitude\ of\ east\ edge\ of\ grid\ cell\ polygon$
S	$:latitude\ of\ south\ edge\ of\ grid\ cell\ polygon$

$$W = lon_f - \frac{gs}{2} \quad (3.5)$$

$$N = lat_f + \frac{gs}{2} \quad (3.6)$$

$$E = W + x_f \quad (3.7)$$

$$S = N - y_f \quad (3.8)$$

$$tile\ polygon = [(W, N), (E, N), (E, S), (W, S)] \quad (3.9)$$

3.2.1.2.2 Standardized Grid Sizes The initial driving force behind the creation of iDEM is to be able to create DEM at various grid resolutions as required.

Ideally, tasks such as accurate tsunami modelling should be conducted using a high resolution DEM. Tsunami modelling is computationally intensive, and the higher the DEM resolution the more computation is required. To reduce the computational overhead researchers use nested DEM of increasing resolution when modelling tsunamis. Lower resolution DEM are used to calculate the boundary conditions of the model with higher resolution. The high resolution DEM is used to accurately determine how far inland the inundation will reach along the coastal area of interest. Such modelling is not necessarily only 3 nested DEM levels. Sometimes there are more or less, depending on the source location, geography, and acceptable computational overhead.

Another consideration of equal importance is that the models produce the best results when the nested DEM resolutions change by a factor between 3 and 4 [77]. This means that if a wave source is located far away from the inundation area of interest, several DEM might be nested to reduce computational requirements while maintaining the step resolution change in the region of 3 to 4 (e.g. 144" - 36" - 9" - 3" - 1" - 1/3" - 1/9" are valid steps).

As a third consideration, an initial use of iDEM will be to produce DEM for modellers to directly compare their model results with each other. For such comparison to make sense the modellers will need to use the same sequence of DEM to produce their model results. The system should also be able to produce DEM at custom resolutions but a set of standard resolutions should comprise the core of the DEM collection.

Selected standard grid sizes. If they are chosen correctly there is a significant advantage to producing DEM using iDEM. Standard resolutions have been chosen so that they all fit exactly within a standard tile size, meaning that a standard tile will only contain whole grid cells at all resolutions. It follows that standard tiles of neighbouring areas can be laid side by side to assemble a larger DEM just like pieces of a puzzle.

In the United States tsunami modelling, inundation analysis and warning of impending tsunami disasters falls under the umbrella of the National Oceanic and Atmospheric Administration (NOAA). Current warnings of imminent inundation issues along the Canadian west coast originate from NOAA. NOAA generates DEM at certain common grid sizes [95]. DEM grid size is inversely proportional to resolution;

the smaller the grid size the higher the resolution. The grid sizes are expressed in terms of minutes or seconds of geographic arc angle. The common grid sizes used by NOAA are:

- 2 arc minutes
- 48 arc seconds
- 36 arc seconds
- 30 arc seconds
- 24 arc seconds
- 6 arc seconds
- 3 arc seconds
- 1 arc second
- 1/3 arc second
- 1/9 arc second

Tsunami modellers often use a 1 or 2 arc minute grid size as the lowest resolution DEM input. Interestingly a subset of the grid sizes above can be multiplied by an integer to total two arc minutes, allowing exact nesting. The step size between resolutions include 2, 3, 4 and 5; not just 3 and 4 .

- 2 arc minutes
- 30 arc seconds
- 6 arc seconds
- 3 arc seconds
- 1 arc second
- 1/3 arc second
- 1/9 arc second

The Global Multi-Resolution Topography synthesis stores 10 layers of DEM tiles of doubling resolution (see Section 2.8.1). This is a logical approach which can be adapted to the standard resolutions to be stored for tsunami modelling. Looking at the four smallest grid sizes it happens that they all divide cleanly into 1° (3600 arc seconds) and that they reduce by a factor of 3 with each size. Replacing the 2 arc minute size with a 144 arc second size introduces a resolution of approximately $2\frac{1}{3}$ arc minutes. Replacing the 30 with a 36 arc second grid and adding a 9 arc second size completes the sequence. The resulting standard sizes alter by a factor of 4 above the 9 arc second grid size and a factor of 3 below. Standard iDEM grid sizes chosen are:

- 144 arc seconds
- 36 arc seconds
- 9 arc seconds
- 3 arc seconds
- 1 arc second
- 1/3 arc second
- 1/9 arc second

This sequence of standard resolutions do not use the same sizing factor. Even though all the resolutions factor cleanly into the standard tile size (1° = 3600 arc second) the

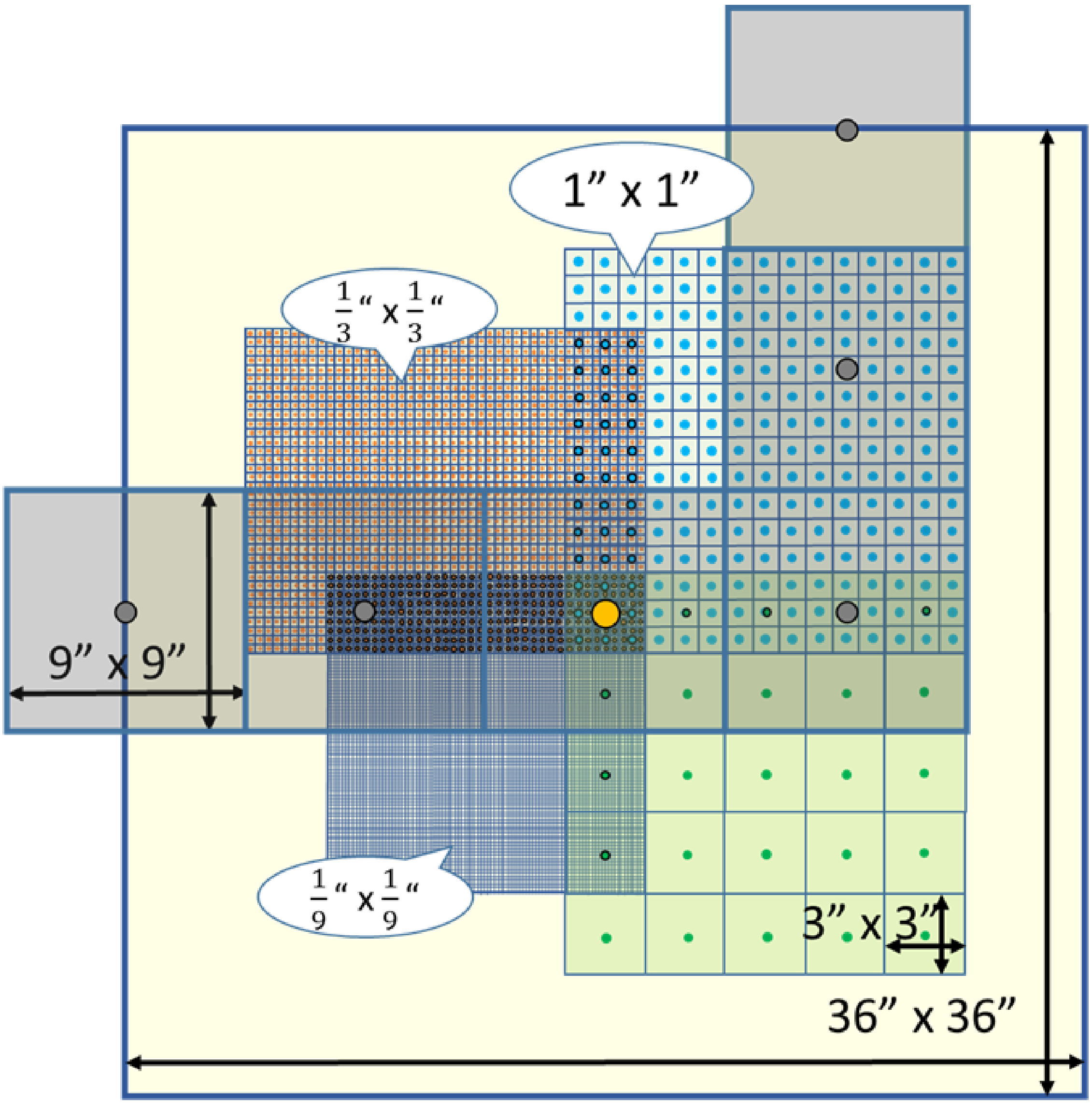


Figure 3.9: Interlocking tile resolutions.

sizing step changes from a factor of 4 to 3. This effectively divides the standard into two sets. The first set consists of 144, 36 and 9 arc seconds, the second set contains all the grid sizes from 9 to 1/9 . The 9 arc second grid size is common to both sets. This implies that in order to maintain standard grid sizes the tile sizing can only occur as integer multiples of 9 arc seconds. This should not have a significant impact on the actual operation of the designed system. A minimum tile size of 9 arc seconds at 50° latitude corresponds to a tile size of approximately $179 \times 278 meters$.

Grid tiles. Effectively merging the datasets requires the data to be analysed in groups defined by the intended DEM grid. In a GIS enabled database a good way of doing this is to define polygons for each grid cell area. The same method can be used to define each cell as was used to define the tile border.

Given that:

$$
\begin{array}{ll}
lon_f & : final\ anchor\ longitude\ (as\ per\ 3.1) \\
lat_f & : final\ anchor\ latitude\ \ \ (as\ per\ 3.2) \\
gs & : grid\ size\ (indecimaldegrees) \\
x_f & : final\ tile\ size\ x\ direction\ (as\ per\ 3.3) \\
y_f & : final\ tile\ size\ y\ direction\ (as\ per\ 3.4) \\
W & : longitude\ of\ west\ edge\ of\ grid\ cell\ polygon \\
N & : latitude\ of\ north\ edge\ of\ grid\ cell\ polygon \\
E & : longitude\ of\ east\ edge\ of\ grid\ cell\ polygon \\
S & : latitude\ of\ south\ edge\ of\ grid\ cell\ polygon \\
GP & : grid\ cell\ polygon
\end{array}
$$

$$
\begin{array}{l}
\textbf{FOR}\ \ i = 0\ \ to\ \ ((x_f/gs) - 1) \\
\quad \textbf{FOR}\ \ j = 0\ \ to\ \ ((y_f/gs) - 1) \\
\qquad W = lon_f - (gs/2) + (gs \times i) \\
\qquad N = lat_f + (gs/2) - (gs \times j) \\
\qquad E = W + gs \\
\qquad S = N - gs \\
\qquad GP_{ij} = [(W, N), (E, N), (E, S), (W, S)] \\
\quad \textbf{ENDFOR} \\
\textbf{ENDFOR}
\end{array}
\tag{3.10}
$$

3.2.2 Stage 2B Select Datasets

3.2.2.1 Select Tile Datasets

If more than one dataset intersect with the tile definition, and have points within the tile, then in order to select which datasets should be used in the creation of the DEM we will need to identify datasets that intersect the tile polygon and, select the datasets to be use in the DEM creation.

3.2.2.2 Select Boundary Datasets

To prevent tile join distortion, tiles are given overlapping boundaries (see Section 2.8.1). During the gridding and interpolation processes that occurs in Stage 3 a trend surface is created. In order to prevent tile edge distortion the trend surface algorithm utilizes data populating 10 grid cells beyond the tile edges. At the same time the CUBE gridding algorithm requires data populating a distance of 5% $of\ depth$ beyond the tile edge to prevent tile edge distortion. Both of these conditions need to be satisfied.

This means that additional datasets within the boundary that are not within the tile need to be selected. In order to do this, we need to:

- determine the boundary size (bs) and the depth. In order to minimize outlier distortion the depth should be determined by fitting all tile data points to a normal distribution and then setting the $depth = \mu - 2\sigma$ (95.45% of all points).

  Given: $\lceil A \rceil \equiv round\ up\ A \qquad gs = grid\ cell\ size$

  $$G_{5\%} = \left\lceil \frac{depth \times 0.05}{gs} \right\rceil \tag{3.11}$$

  where $G_{5\%} = number\ of\ bounary\ grid\ cells\ CUBE\ requires$

  $$bs = greater\ of\ G_{5\%}\ and\ 10 \tag{3.12}$$

  where $bs = boundary\ size\ measured\ in\ number\ of\ grid\ cells$

- identify datasets that intersect the boundary polygon that have not already been selected as intersecting with the tile polygon

- select datasets to use in DEM creation

3.2.2.3 Define Boundary Grid

Use the method detailed in Section 3.2.1.2.2 Defining Tile Grid to define each grid cell.

3.2.3 Stage 2C Merge Datasets

Merging the datasets selected within the tile and its boundary requires no special processing if all the datasets are isolated from each other. On the other hand if two or more datasets overlap distortions can be introduced into the resulting DEM. The following discussion provides a simple method for merging overlapping datasets and having some control over resolution distortion. This method is included here as a simple option. By integrating additional merging modules more sophisticated algorithms can replace this simple merging algorithm.

3.2.3.1 Resolution Distortion

Resolution distortion can occur in regions where datasets overlap. Common resolution distortions can be seen in the simplified diagrams in Figure 3.10. The diagrams are grouped into two rows of three panels.

The first row shows two datasets of similar resolution merged by two different methods. The *by-grid-cell* method (second panel first row) shows that the merged datasets produce a good reproduction of the terrain. The *by-hull* method (third panel first row) shows that the merged datasets produce a distorted reproduction of the terrain. In the case of merging similar resolutions the by-the-grid method is superior.

The second row shows two datasets of dissimilar resolution merged by two different methods. To clearly show this to be a resolution distortion one dataset is actually a low resolution version of the other. The low resolution dataset is represented by a single yellow elevation point. The purple dataset has a resolution that is 15 times greater. Both datasets represent the same terrain. The lower resolution dataset is actually the average elevation value of the higher dataset.

The by-grid-cell method (second panel second row) shows that the merged datasets produce a distorted reproduction of the terrain. The by-hull method (third panel second row) shows that the merged datasets produce a good reproduction of the terrain. In the case of merging dissimilar resolutions the by-hull method is superior.

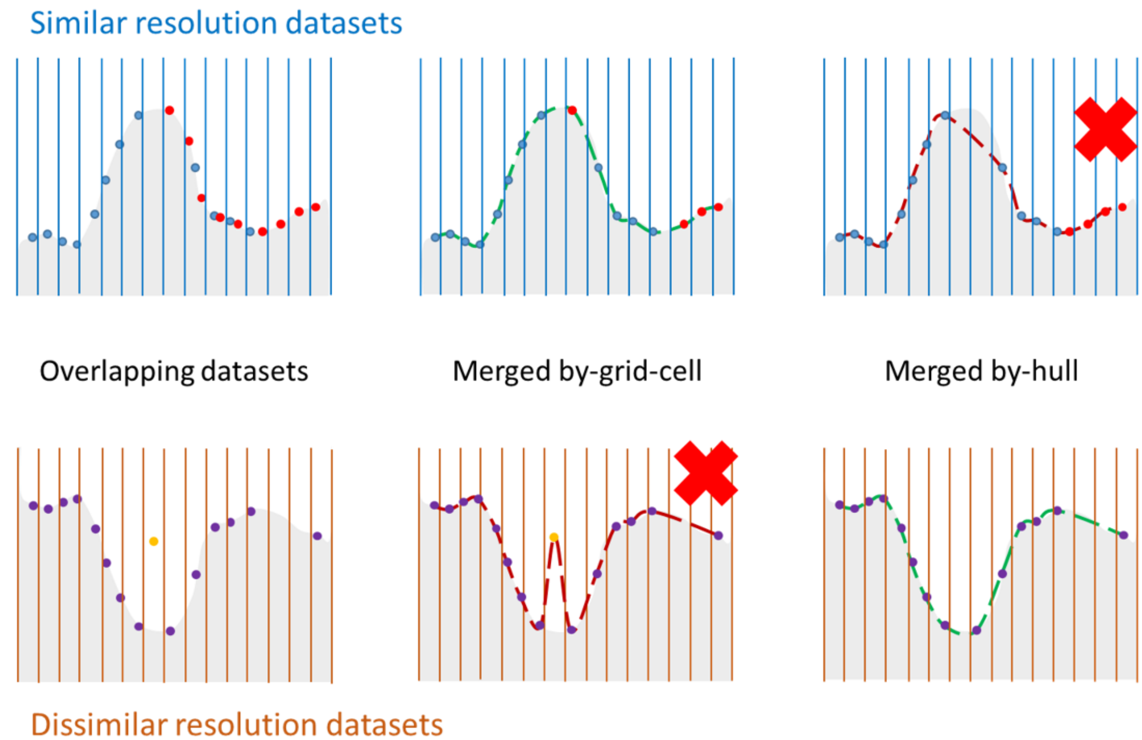


Figure 3.10: Simplified examples of resolution distortion (merging methods examined in Sections 3.2.3.3 and 3.2.3.4).

These two merging methods are complimentary; by-grid-cell is better for similar datasets and by-hull for dissimilar datasets.

3.2.3.2 Sorting Methods.

When two overlapping datasets (see Figure 3.11) are merged an algorithm is required to sort which points to keep and which to discard. This algorithm is the sorting method.
Sorting methods:

- Keep All Points
- Highest *Density*
- Lowest *Uncertainty*
- Highest $\left(\frac{Density}{Uncertainty}\right)$

(3.13)

The first sorting option is trivial, as no merging process needs to be applied.

When sorting by highest *Density* it is important to remember that a portion of the input data is in the form of existing DEM. It is possible that an overlapping single beam or even a lead-line survey has a higher density than those DEM. If the overlapping sets are weighted by density the DEM data could be decimated or even eliminated in order to keep the denser vintage data. This would probably not be a desirable outcome.

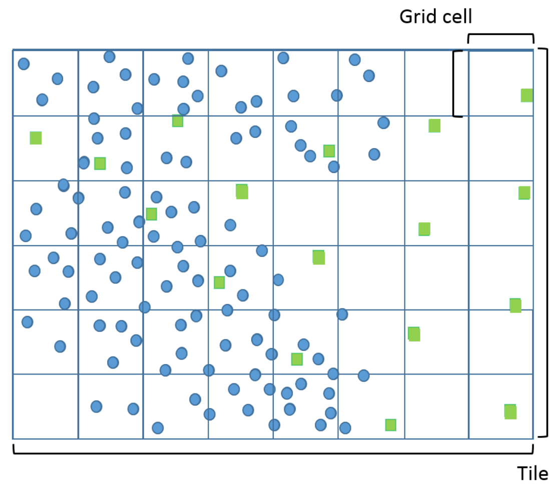


Figure 3.11: Two overlappig datasets.

In many cases DEM grids have good *Uncertainty* values. If overlapping datasets are weighted by uncertainty it is likely that DEM grids will tend to eliminate vintage datasets. This result is not guaranteed, as a vintage dataset could have low estimates of uncertainty.

A more robust comparison would be to compare the factor $(Density/Uncertainty)$. It is less likely that a vintage dataset would be both denser and have low enough uncertainty to be selected over a DEM.

3.2.3.3 Merging Method: by-hull.

In some geographic information systems (GIS) the term hull is used to refer to a polygon that encases a dataset.

If the overlapping datasets are sorted using the $(Density/Uncertainty)$ rule then for the data points of the datasets within the tile, let average $Uncertainty = 5$ for round, blue, dataset and average $Uncertainty = 4$ for square, green, dataset (see Figure 3.12).

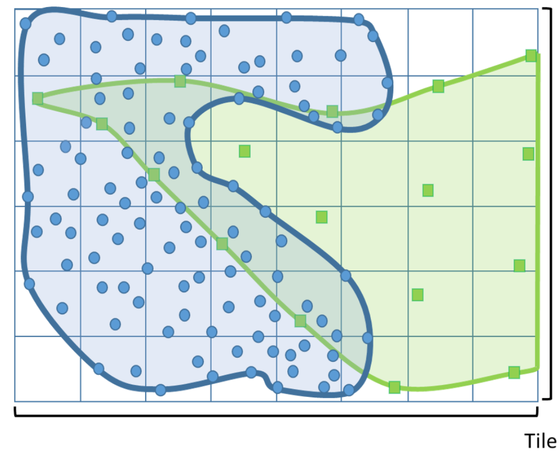


Figure 3.12: Two example datasets with hulls.

$$Resolution \propto Density = \left(\frac{dataset\ point\ count}{hull\ area}\right) \tag{3.14}$$

The result is shown in Figure 3.13. The discarded points are shown as smaller red squares. They all originated from the square, green, dataset. Note that the two datasets are now separate and distinct, and there is no mixing of datasets. Some cells may lose all of their data, as highlighted in the circled cell.

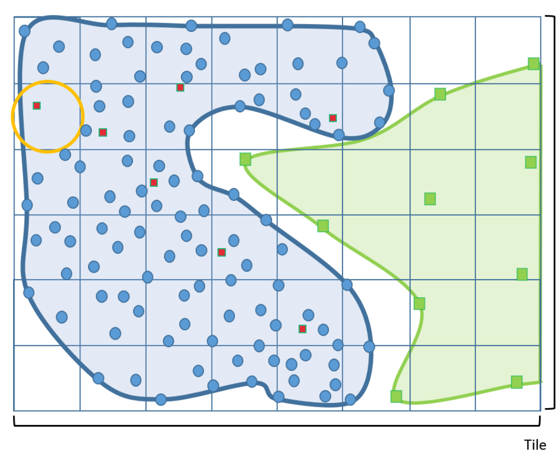


Figure 3.13: Two example datasets merged by-hull.

3.2.3.4 Merging Method: by-grid-cell

If the overlapping datasets are sorted using the same ($Density/Uncertainty$) rule with the same average dataset uncertainties but merged using the by-grid-cell method a different result is obtained (see Figure 3.14).

When using the by-grid-cell merging method the definitions of $Density$ and $Uncertainty$ change.

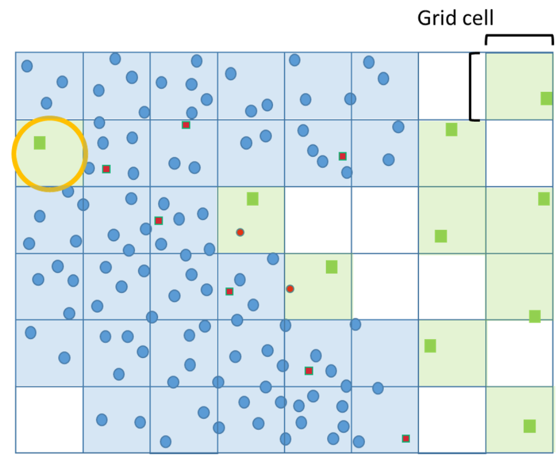


Figure 3.14: Example of two datasets merged by-grid-cell.

$Uncertainty = (average\ uncertainty\ of\ the\ points\ of\ a\ dataset\ within\ a\ grid\ cell)$

$$Density = \left(\frac{number\ of\ points\ of\ a\ dataset\ within\ a\ grid\ cell}{area\ of\ a\ grid\ cell}\right) \tag{3.15}$$

Since the area of every grid cell is identical:

$$Density \equiv (number\ of\ points\ of\ a\ dataset\ within\ a\ grid\ cell) \tag{3.16}$$

In this example the uncertainties are kept the same as for the by-hull example to eliminate uncertainty as a factor in how the two merging methods differ. For most datasets each point will have an uncertainty value, and it is unlikely that the average uncertainty of the dataset within the tile would be identical to the average uncertainty of the dataset within each grid cell.

Using the by-grid-cell merging method allows the datasets to intermingle and ensures that the maximum number of grid cells are populated. This is highlighted by the circled data point that was eliminated when using the by-hull merging method, as it remains as a valid point.

3.2.3.5 The Merging Process.

Utilizing the complementary merging methods in iDEM requires that the method selection be automated. The sorting method will be an input parameter when defining the tile. Prior to actually merging the datasets, they need to be adjusted to the mean sea level (MSL) tidal datum.

If the available tidal models do **not** include the tile area and the tile is adjoining one or more areas with tidal models, then the tidal model is based on the adjoining models. If the tile is isolated from existing tidal models the system is designed to act in two ways depending on whether an override is set.

1. If the override is not set the process will stop.

2. If the override is set the program will assume the missing datum is MSL and increase point uncertainty of all points by half the maximum tide variation of the neatest available tidal model.

If a tidal model exists for the tile area and the dataset datum is known then the program will adjust to MSL.

If a tidal model exists for the tile area but the dataset datum is **not** known and the dataset overlaps, or is adjoining to, a dataset(s) with known datum the program will infer the unknown datum and adjust to MSL. If the dataset is isolated from the existing known datum the system is designed to act in two ways depending on whether an override is set.

1. If the override is not set then the program will deselect the dataset.

2. If the override is set the program will assume that the datum is MSL.

Once the data has been referenced to the same datum, the datasets are ready to be merged. In order to enable iDEM to merge datasets while managing resolution distortion a simple automated algorithm was developed as part of this research. The algorithm automatically selects the appropriate merging method. The complementary nature of the two merging methods is based on the ratio of the resolutions of the datasets. A threshold is used to classify the datasets as having similar or dissimilar resolutions.

The threshold can be referred to as a *similarity threshold* (ST).

$$\textbf{IF} \quad ST \quad \leq \quad \frac{|ResolutionDataset_A - ResolutionDataset_B|}{ResolutionDataset_A + ResolutionDataset_B} \quad \textbf{THEN}$$

$$merge \;\; overlapping \;\; datasets \;\; using \;\; the \;\; by - grid - cell \;\; method$$

$$\textbf{ELSE}$$

$$merge \;\; overlapping \;\; datasets \;\; using \;\; the \;\; by - hull \;\; method$$

$$\textbf{ENDIF} \tag{3.17}$$

Note: $0 \leq ST \leq 1$. ST should be a rational number between 0 and 1. If ST is set to a value of 0 then all merging will exclusively be by-grid-cell method; if set to a value of 1 then all merging will exclusively be the by-hull method.

This automated merging algorithm allows the user to define the similarity threshold to control the level of acceptable distortion. A default value of $ST = \frac{1}{2}$ will be utilized in the prototype of iDEM. A qualitative analysis suggests that this is an acceptable compromise between minimizing data loss and minimizing resolution distortion. Quantitative analysis of the resulting DEM will be required to ensure that this is an acceptable default value for an inexperienced user to rely upon.

3.3 Stage 3 - Grid and Interpolate

An examination of Gridding Methods (see Section 2.7) and Some Existing Fusion Methods (see Section 2.8) in light of the need for elevation with uncertainty data to be transformed into DEM with DCM leads to the following conclusions:

- The Three Step Approach (3Step)
  - fullfills the requirement to create DEM with DCM;
  - is a fusion process that can deal with any density of data ; and
  - is a process incorporating some computationally intensive components that may make processing excessive amounts of high density data time prohibitive.
- The Combined Uncertainty and Bathymetry Estimator (CUBE)
  - considered by many to be the current best practice for high density bathymetric data processing;
  - fullfills the requirement to create DEM with DCM for bathymetric data;
  - discards low density bathymetric input data;
  - unable to process topographic input data; and
  - does not interpolate so a grid cell with insufficient high density bathymetric data will be left blank.

In light of these facts:

- high density bathymetric portions of the merged dataset can be gridded by CUBE into low density DEM;
- the new low density DEM can be combined with the original merged dataset in such a way as to reduce data density to create a low density modified merged dataset; and
- the modified dataset can be gridded and interpolated by the 3Step process.

This covers all data combinations except high density topographic input data. Allowance should be made for future integration of the ability to grid high density topographic input data. As with CUBE such a module would create a low density

DEM that can be integrated into a modified merged dataset to produce a low density DEM. At this time it is assumed that high density topographic data (e.g. Lidar data) will be preprocessed into DEM prior to being imported into iDEM.

The iDEM system's third stage, grid and interpolate, is an amalgamation of CUBE and 3Step to produce DEM with DCM.

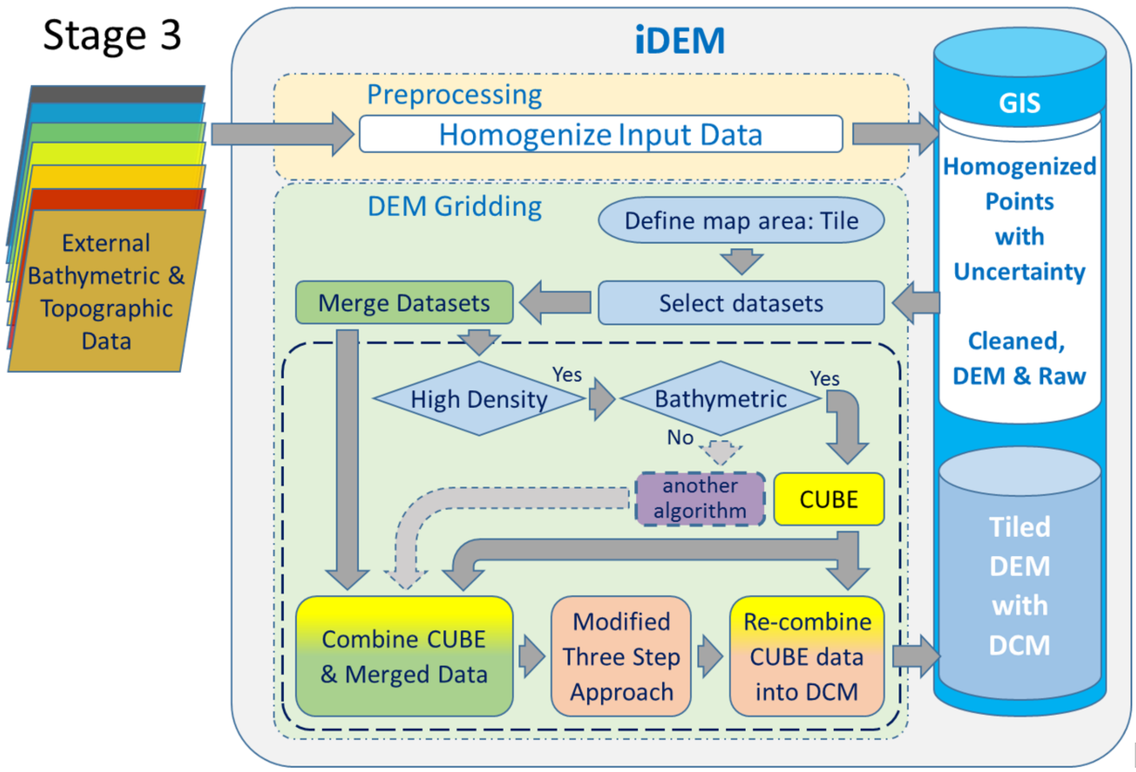


Figure 3.15: iDEM Stage 3 - Grid and Interpolate detail.

The Stage 3 overview shown in Figure 3.15 can best be appreciated as three gridding and interpolation steps. These steps will be examined from the viewpoint of how they process the datasets from Stage 2:

Step A) Grid high density data

Step B) Interpolate low density data

Step C) Produce final product

3.3.1 Stage 3A Grid High Density Data

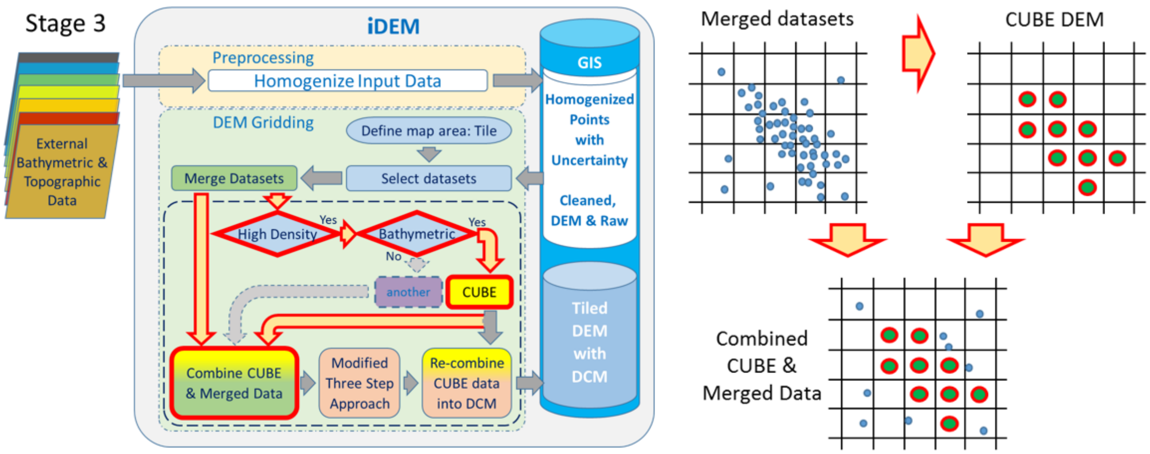


Figure 3.16: Combine CUBE and merged data.

High density bathymetric data that is processed with CUBE will have associated uncertainty values. The *CUBE DEM* shows the output when the input data density threshold is set to a minimum of five data points (high density data can be considered to be at least five data points per grid cell [115]; although CUBE can function with as little as one data point per grid cell this compromises accuracy). The elevation value is represented by the center of each point (green) and the uncertainty by the (red) ring surrounding each point.

CUBE's output is gridded to the desired cell size and it has significantly lower density than the input data. It is clearly evident that where the CUBE input data did not meet the required threshold the grid cells are empty.

CUBE's output is combined with the original merged datasets and the resulting *Combined CUBE & Merged Data* will be input for a Modified Three Step Approach.

3.3.2 Stage 3B Interpolate Low Density Data

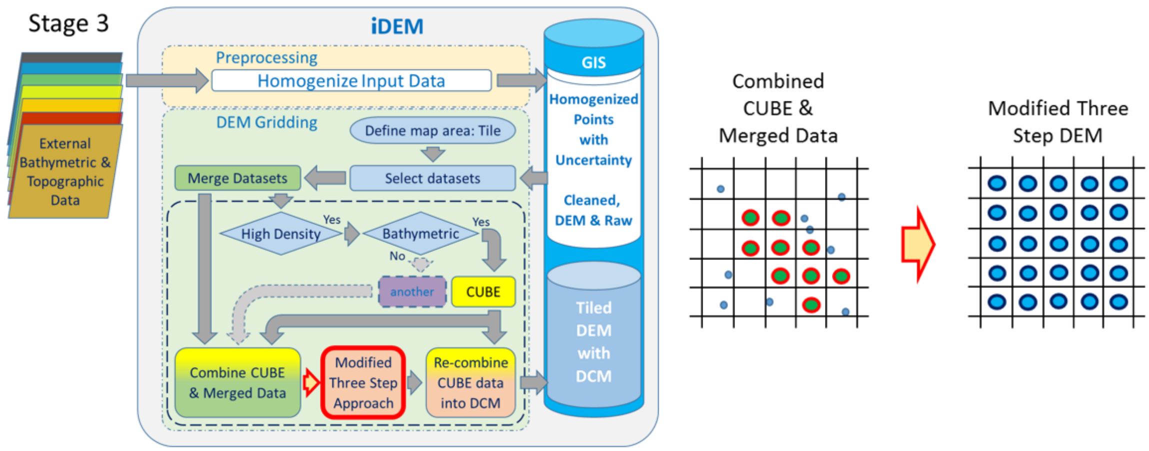


Figure 3.17: Modified Three Step Approach.

The *Combined CUBE & Merged Data* is now processed by a Modified Three Step Approach (M3Step). The *Modified Three Step DEM* output is a fully populated DEM with uncertainty.

3.3.2.1 Modified Three Step Approach

The M3Step is based on the 3Step shown in Section 2.8.3. Two changes have been made to the original process design:

- iDEM homogenization ensures that all data points have an uncertainty value; the Monte Carlo uncertainty estimation shown in the original 3Step (see Figure 2.42 on page 55) resides in Stage 1 of iDEM rather than in the M3Step which is in Stage 3.
- ordinary Kriging is subject to aliasing, so an optional anti-aliasing filter has been introduced to reduce distortion (see Figure 3.18).

3.3.2.1.1 Reducing Sampling Distortion. *CUBE DEM* does not contribute to aliasing in the final grid interpolation. Lower density portions of the *Merged Datasets* may contribute to sampling distortion, especially in overlapping regions if points from more than one dataset are kept. The potential for sampling distortion increases if there is more than one point value in a given grid cell. This effect can be minimized by aggregating the points in the cell.

CUBE automatically aggregates multiple points in a cell with a sophisticated algorithm with weighting related to uncertainty. An low pass filter can be implemented for the M3Step by aggregating multiple points in a grid cell and taking the mean of the point locations [121]. The filter's place in the process ensures that the variogram estimation is the same regardless of whether the filter is utilized or not.

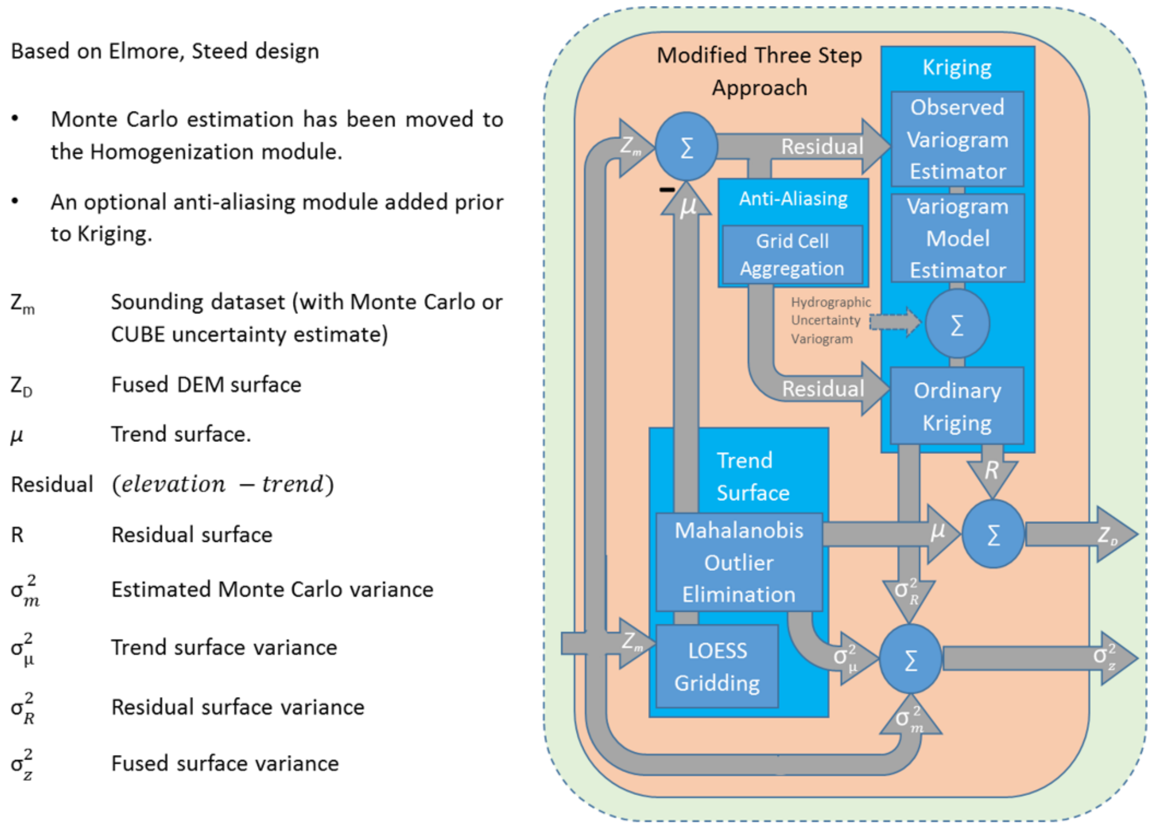


Figure 3.18: Modified Three Step Approach detail.

3.3.2.1.2 Outlier Detection is implemented in the original design of 3Step as part of the *Trend Surface* (see Figure 2.42). A Mahlanobis unit is one standard deviation of a normal distribution. In the 3Step design an outlier is defined as a data point if a depth estimate on the trend surface differs from the mean of the surrounding area by more than three Mahalanobis units; that point is marked as no data. The M3Step includes the same provision, with the outlier detection module being external to the main code. This allows the algorithm utilized for outlier detection to be altered, enhanced, or replaced without requiring major code revision. For example, the simplest adjustment might involve altering the filtering distance. Any multiple of Mahalanobis units, or fraction thereof, can be utilized for filtering data for outliers.

3.3.3 Stage 3C Produce Final Product

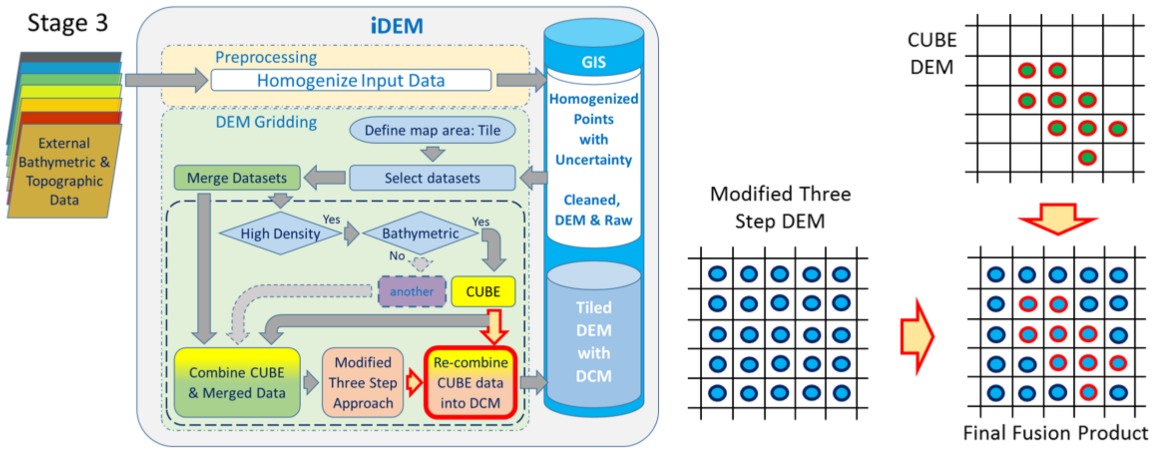


Figure 3.19: Re-combine CUBE and Modified Three Step DEM.

When an algorithm processes data the uncertainty of the output is generally larger than the uncertainty of the input. The M3Step will not alter the elevation values of the CUBE portion of the input data but it will increase the uncertainty of those values. The CUBE uncertainty values are more accurate and the M3Step can be considered to have artificially inflated those values. The accurate CUBE uncertainty values need to be restored to the DCM to create the *Final Product* (see Figure 3.19).

3.4 DEM to Order

The design calls for the *Final Product* DEM to be stored in iDEM. Each stored DEM will have a DCM and a metadata history detailing the datasets and process used to create them.

Multiple DEM may cover the same general area of interest at differing resolutions (grid sizes). Ideally all of these DEM will have a tidal datum of mean sea level.

When a DEM is going to be packaged for delivery it should first be adjusted to the desired tidal datum. The details of this adjustment should be appended to the metadata file. The DEM and DCM need to be exported in the desired file format.

The metadata file will initially be exported as a CSV file. CSV files are an ASCII text based format that are easy to access and are well understood. Other formats can be added to the system at a later stage if required.

Chapter 4

Implementation

Many of the algorithms utilized in the iDEM system are implemented in other existing applications. Rather than re-implement those algorithms iDEM has employed algorithms already implemented by integrating existing applications into a new framework. The iDEM system does not seek to fully integrate the existing applications by accessing all of their functionality; but rather to mimic and automate the actions of a user who utilizes small portions of the various applications manually to achieve the same overall process. The initial case study will automate the process to produce DEM to be used in tsunami modelling.

4.1 Software

4.1.1 A Controlling Framework

In order to utilize existing application algorithms a suite of suitable applications were selected. The iDEM system is developed as a python library that constitutes a controlling framework. This automated control script can be used from the command line or can be integrated into friendly general user interfaces (GUI).

One algorithm that has not been available in an existing application is the automation of the dataset merging. This algorithm is designed as a separate python application that is itself integrated into the overall framework.

All major interface modules like the uncertainty estimation, trend surface, Mahalanobis outlier detection and residual surface calculation, should be implemented outside the framework. The framework will continue to function regardless of which of the available algorithms is integrated to perform a particular task. This means that each iDEM module requires a python wrapper to interface seamlessly with the framework. Such a structure will enable iDEM to be enhanced, upgraded or altered in a flexible manner that will minimize any required re-coding of the actual framework. A complete application integration can be replaced with the only requirement being to rewrite the application wrapper.

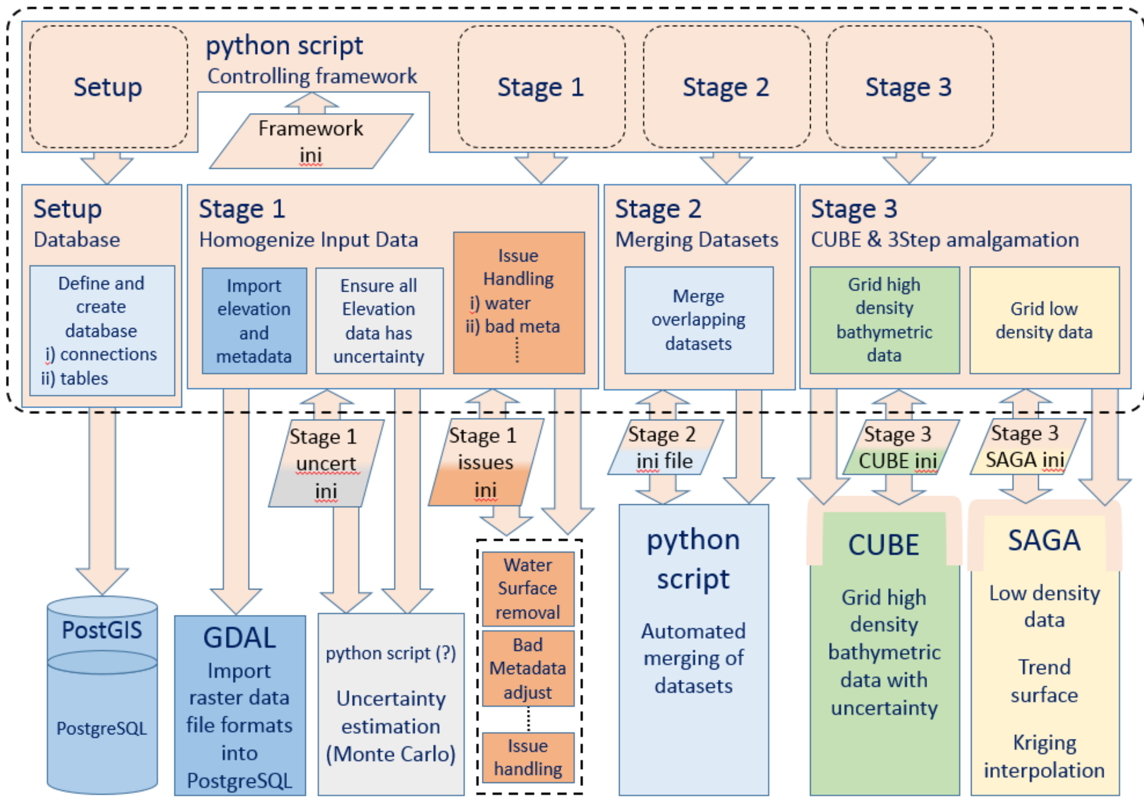


Figure 4.1: iDEM control framework: major components.

The one core application that does not fall into this category is the relational database. All data flow to and from the database so all other integrated applications need to interface with it directly or indirectly. Changing this core component would require rewriting all major module wrappers and the framework itself. For this reason the database and the framework are considered as a single unit.

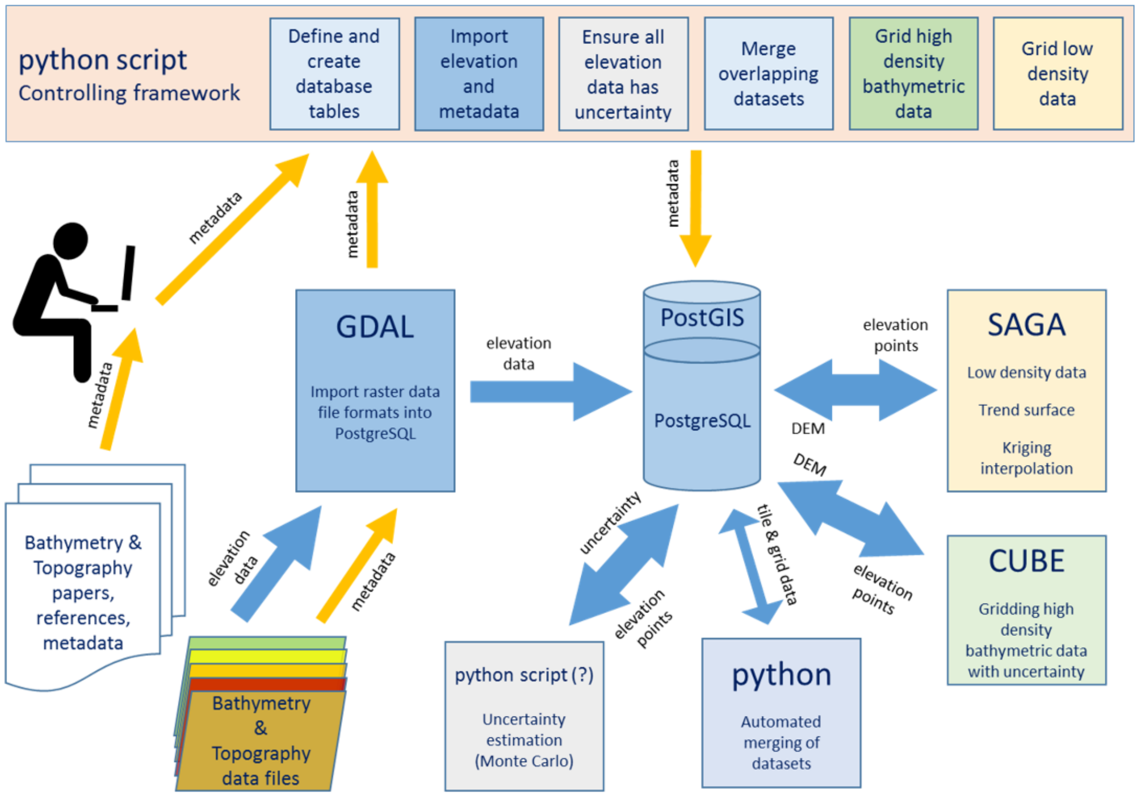


Figure 4.2: iDEM control framework: data flow.

4.1.2 OpenSource

OpenSource applications have been favoured in the search for suitable candidates. The main reasons for this is:

- minimal capital outlay;
- flexible licence agreements; and
- greater choice of operating system requirements.

In making this choice consideration was also given to the fact that OpenSource applications also have some drawbacks:

- maintenance can cease at any time, proprietary application maintenance is often driven by economic factors that promote continuity and stability;
- documentation can be inadequate since the main users are often the developers; and

- technical assistance is usually lacking due a lack of interest and/or funding.

The majority of the commercially available proprietary applications that could be suitable for iDEM run on the Windows operating system. The majority of Open-Source applications can run on both Windows and Linux based operating systems.

4.1.3 Software Choices

In order to keep iDEM flexible Linux has been selected as the primary operating system and Windows as the secondary.
As a programming language python script has been selected because it is known for its integration flexibility. It has been used for the framework of the controlling software. All the selected components are accessible from python script and it is a known for effective systems integration [107].
The following OpenSource applications have been integrated into iDEM:

- PostgreSQL: This proven enterprise class relational database was selected as the core of iDEM. This is an object oriented application. With more than 15 years of active development, PostgreSQL is a proven platform with programming interfaces that include python and ODBC. The deciding factor to use this application is that it fully supports geographic information system functionality with the installation of PostGIS [105].

- PostGIS: This database extender for PostgreSQL enhances the PostgreSQL database with spatial functionality. Spatial functionality provides support for geographic objects, allowing location queries to be run in SQL. PostGIS offers not only basic location awareness but also extra functionality along with added types such as geometry, geography and raster. It is known to be a fast, feature rich, and reliable spatial database management system [104].

- GDAL: Geospatial Data Abstraction Library is a translator for raster and vector geospatial data formats. This library is available in many programming languages including python. There are also command line utilities for data processing as well as data format translation [49, 124].

- SAGA: System for Automated Geoscientific Analyses software has been chosen to create the trend surface and Kriging processing. These are the two core functions of the Modified Three Step Approach [40] (see Figure 3.18). APIs exist to integrate control from python script and ODBC connection allows access to PostgreSQL databases [32].

Only one proprietary application, CUBE (Combined Uncertainty and Bathymetry Estimator), has been used for the implementation of iDEM.
CUBE has been chosen to process high density bathymetric data. This system greatly assists with data outlier reduction and produces a DEM that includes the high density grid cells but does not interpolate low density grid cells nor fills in empty cells [18, 19]. The *CUBE DEM* output for high density areas is combined with merged data (see 3.3.1. The resulting combined dataset requires interpolation to create the final DEM (see 3.3.2).

4.1.4 Database Tables

The core of the system is the database upon which the python modules and the subsystems they control act. Several tables have been implemented for Stage 1 of the prototype. Many of the data homogenization tables are utilized in subsequent data processing. Temporary tables are utilized during the importation of raster files. One table is created specifically to accept data imported by the GDAL application, a second table is then created to store the data after it is transformed from raster to vector format and adjusted to the default position datum. As part of this process all blank raster data is removed from the temporary vector table. Once the transformed, adjusted and filtered data has been transferred to the permanent data point storage table and any available uncertainty information has been added to the uncertainty tables, these temporary tables are deleted. Details of the permanent tables can be found in appendix A.4.

The first table listed in Appendix A.4 is the *datapoints* table. This table can be considered as the heart of the Data Homogenization stage of the system. This table defines the location of an elevation value as a PostGIS *point* variable type and the actual elevation value as a *double precision* variable type, i.e. as two separate fields. During development it became apparent that it might be better to use the PostGIS spatial data type *pointM* instead of *point*. The *pointM* variable type allows the measurement, or value, at a location to be embedded in the *pointM* variable along with the point's location [104].

Rather than spending time correcting the decision to use *point* in the *datapoints* table the current plan is to transform the *point* and *double precision* elevation fields in

data type	variable holds
point	(longitude, latitude)*
pointM	(longitude, latitude, elevation)

*: When using this type of variable the elevation value needs to be stored in a separate field

Table 4.1: PostGIS data types point and pointM.

the existing datapoints table into a *pointM* field in the tiledata table. This will leverage the additional functionality of the *pointM* data type for DEM creation without requiring time to redevelop the existing code used to populate the *datapoints* table. The tiledata table is part of Stage 2 *Merge Datasets*.

4.1.5 python Library Structure

The current incarnation of the prototype references the acronym oncbf (Ocean Networks Canada Bathymetric Fusion), this acronym has been superseded by the new name iDEM. For consistency with the existing state of the prototype oncbf appears in this document.

4.1.5.1 Command Line Utility - oncbf.py

The system is comprised of several files. A python script oncbf.py is designed as a utility that processes command line options. It reads the initialization file oncbf.ini. Then oncbf.py calls on the *tablesetup* module in the liboncbf.py library to confirm the existence of the required database tables. If any of the required tables do not exist they are created. After ensuring that the command line options appear to be correctly utilized oncbf.py calls the requested module of the library passing it information on how to proceed (see figure 4.3).

4.1.5.2 python Library - liboncbf.py

An overview of the python library is provided in figure 4.4. At this point the functioning modules have been utilized to import a little over 300 million elevation points.

Imported data is automatically processed for known issues such as corrupt uncertainty layers and ocean surface removal. Incoming data is also transformed to the default coordinate system and position datum. When the tidal module is im-

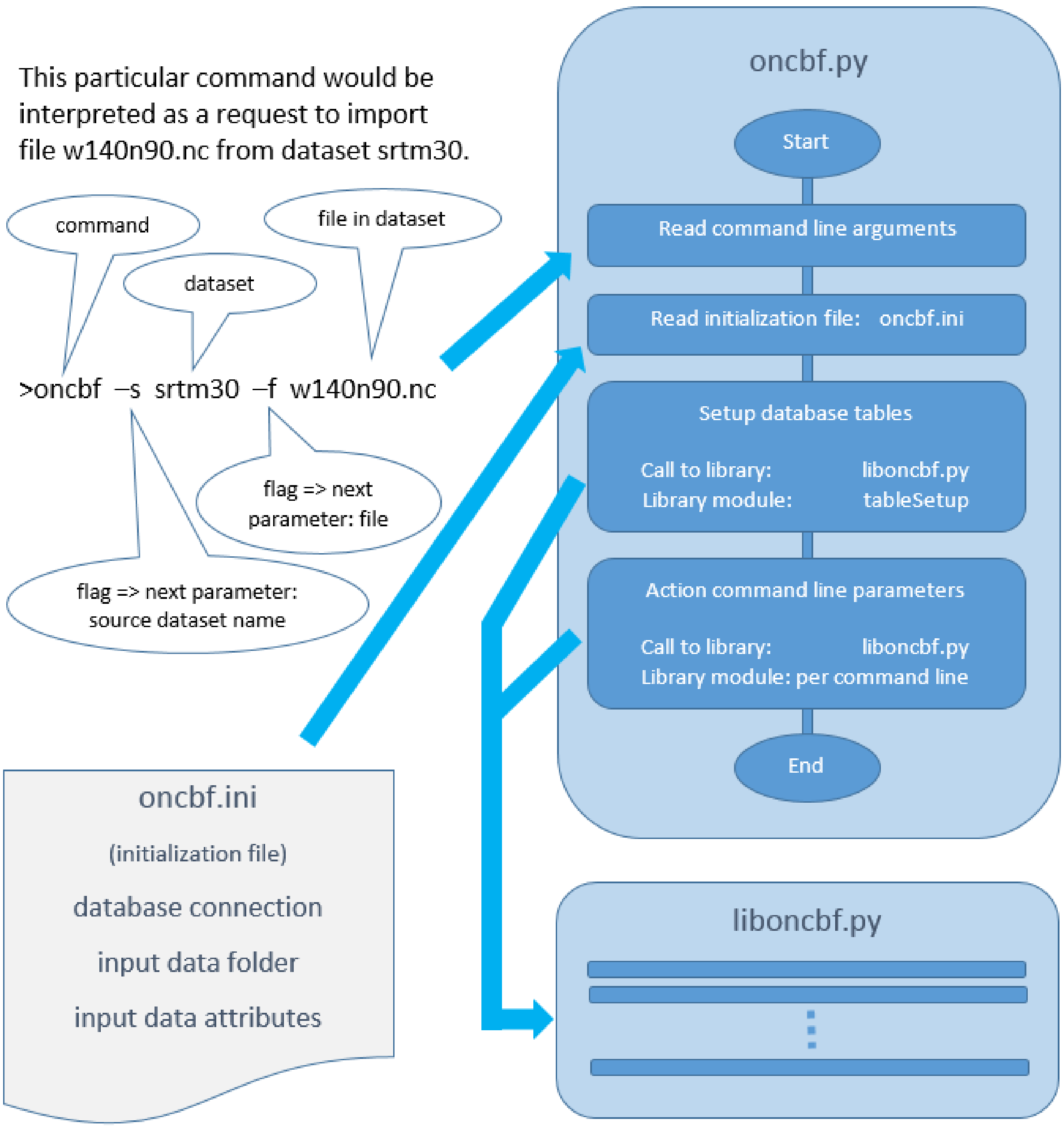


Figure 4.3: Example command line outlining utility functionality: oncbf.py.

plemented the incoming data will not be adjusted to a common tidal datum. Tidal datum adjustment will increase the uncertainty of the data. Since datasets are imported with a particular datum this information will be preserved in its highest state of accuracy. Tidal datum will be adjusted in the merge datasets process of Stage 2.

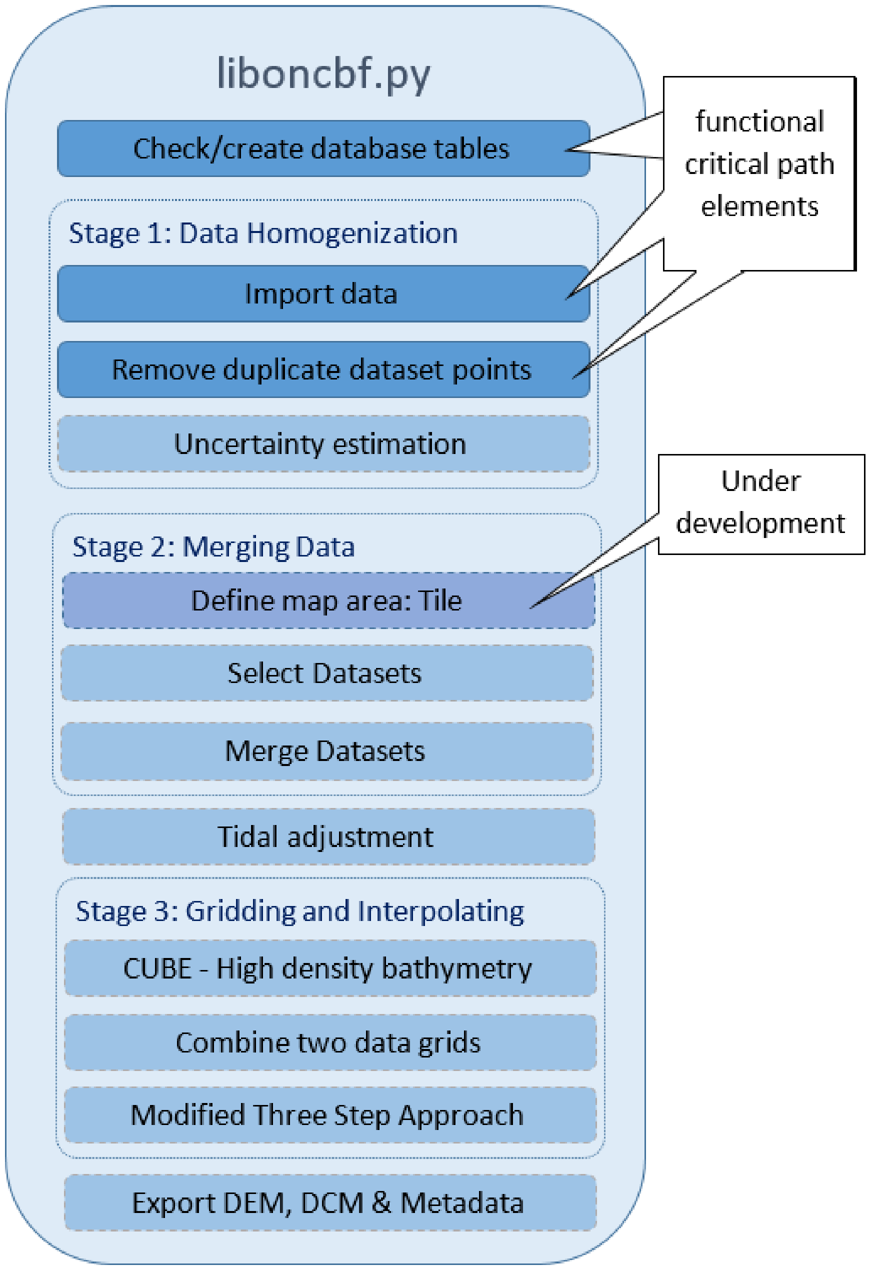


Figure 4.4: Representaive library structure: liboncbf.py.

Some datasets are composed of multiple files (for example the SRTM datasets). In such multi-file datasets each file usually defines a particular area of the Earth's surface. The areas defined in different files often overlap at the boundaries. The SRTM series overlaps by one raster pixel on all edges. To prevent artifacts from forming at these doubled data points a module was implemented to remove duplicated points.

4.1.6 GUI : python library interface

Once the python library is fully functional it is intended that a GUI will be implemented to simplify interaction with and the use of iDEM. Planning for this endeavour has included the use of initialization files for the python library. The initialization files give the library significant flexibility by allowing many settings to be adjusted in simple text files. These files can be easily edited by a GUI before it calls the command line utility to perform an action.

4.1.7 Initialization File: oncbf.ini

At present the initialization file oncbf.ini serves a dual purpose in that it contains the database connection information as well as information about the input data. The final initialization files will be clearly defined and separated based on their function (see Appendix A.3).

Connection information, especially passwords, should be secured in a separate encrypted file or entered by the user when setting up a job to process on iDEM. The password should be erased when the processing has been completed since the database connection will not be required until the next job is undertaken. Security for the framework will need to be addressed in detail as part of future development.

Input data that has not been defined in this configuration file cannot be imported using the library. The user adds the new dataset definition to an initialization file. At present this is accomplished by editing a text file, a GUI could easily simplify this step by guiding the user through the process. The system is not designed to automatically identify data issues at the importing step. An informed user familiar with the data to be imported is required to ensure that potentially crippling issues are identified and dealt with. The need for this definition of the data encourages the user to become familiar with the data prior to importation by defining:

- where input data is located (folder name and location on hard disk),
- which dataset it belongs to (e.g. srtm, CHS500, etc),
- the type or format of the data (e.g. raster, vector, etc; only raster is currently implemented),
- average dataset uncertainty values,

- any known issues with the dataset that may need additional preprocessing (these are flags, in the form of key words, that appear in an initialization text file which tell the system how to customize the data importation).

This requirement also adds flexibility by allowing datasets to be stored in various folders on various drives. It also allows the processing of separate files in a dataset to be split off into different datasets. The processing of different datasets can be partially customized to deal with known issues (new issue handling functions can be added to the python script with system updates).

Aside from added flexibility, the need to set up the initialization files will also reduce the possibility of accidental importation of corrupt or inappropriate data by forcing the user to investigate the data before proceeding.

As the prototype matures this configuration file will be broken into two separate files. The iDEMdb.ini file will contain the database connection information and iDEMdata.ini will hold the input data information. The reason behind this adjustment is to separate two distinctly different types of information. Such separation of function will also make it easier to implement a layered security strategy.

The following is an excerpt from the current oncbf.ini defining the database connection parameters:

[dbTarget]

dbName = onc
userID = user
pwd = password (this line exists for prototyping only)
host = localhost
port = 5432
schema = dev

The next excerpt is from oncbf.ini defining the srtm30 dataset. A line has also been included from 'dummy' dataset to demonstrate the use of the flag indicating corrupt uncertainty layer in a BAG file. Explanations have been added (indented) and some lines shortened for clarity:

[data]

srtm30Folder = data/srtm30/srtm30/grd/
    defines the folder in which the dataset resides
srtm30Extension = .nc
    specifies the extension of the dataset files
srtm30Type = raster
    specifies the general format datafile
srtm30MultiFile = True
    specifies that data files in the dataset overlap
srtm30dsUref1 = Rodriguez, Ernesto, ... the SRTM topographic products." (2005).
    metadata reference for dataset's average uncertainty figures [111]
srtm30dsUx1 = 12.6
srtm30dsUxUnit1 = meter
srtm30dsUy1 = 12.6
srtm30dsUyUnit1 = meter
srtm30dsUz1 = 9
srtm30dsUzUnit1 = meter
    details dataset's average uncertainty and units (meters or percent)
srtm30dsUref2 = Wagner, M. "SRTM DTED Format." Product Desc...sion 1 (2003).
    2^{nd} metadata reference for dataset's average uncertainty figures [122]
srtm30dsUx2 = 20
srtm30dsUxUnit2 = meter
srtm30dsUy2 = 20
srtm30dsUyUnit2 = meter
srtm30dsUz2 = 16
srtm30dsUzUnit2 = meter
    dataset's 2^{nd} average uncertainty and units (meters or percent)
srtm30SRIDcode = 4326
    srtm30 metadata doesn't explicitly include SRID code, this provides the missing information manually

dummybagissue1 = uncertcorupt
setting the flag for corrupt uncertainty layer

For a full printout of oncbf.ini see appendix A.5.1.

4.1.8 Command Line Options

When properly registered within the operating system oncbf.py can be executed as a command line function. The following is a partial prototype with Stage 1 options. This version of the function has several options or command line parameters that can be used to determine the actions performed by the underlying library liboncby.py.

The first two parameters need to be used together. The data file to be loaded must always be referenced to a data source as initialized in oncbf.ini (see Initialization file section 4.1.7).

parameter		
short	long	description/help
-f	–datafile	Input data file name
-s	–datasource	Input data file source
-R	–reload	Set to True / False to reload existing data
-K	–clean	Set to True / False to do slow bulk clean of data points to remove duplicate rows within datasets

Table 4.2: Example command line parameters.

The following command will attempt to load datafile w140n90.nc from dataset srtm30. If it is already loaded this command will return a message to that effect.

oncbf -s srtm30 -f w140n90.nc

It is possible to reload data that is already in the database. Appending the -R parameter instructs the library to delete the existing data from this file that is already in the database and then load the file again.

oncbf -s srtm30 -f w140n90.nc -R True

The library ensures that any one dataset that consists of multiple files that have overlapping elements is prevented from storing its duplicate data in the database.

This is accomplished by first loading a new file from the dataset and then searching for duplicated points within the loaded dataset and deleting them. This will occur automatically if the line *srtm30MultiFile = True* is set in the initialization file (see configuration files section 4.1.7 and appendix A.5.1). If this initialization line has been overlooked and there is a suspicion that duplicated dataset points are stored in the database it is possible to clean the database. This is a slow process as every dataset in the database is searched for duplicate points which can take many hours of processing depending on the size of database. The following command initiates the cleaning process:

oncbf -K True

Additional options will be added as the prototype matures. The parameters will allow each stage of iDEM to be executed separately. Control of individual functions of each stage will also be available to allow knowledgeable users to adapt or modify the process.

4.2 Hardware Considerations

iDEM and the chosen component applications can run on the average desktop or notebook computer with some resource constraints. If the input data and consequent output is confined to a local area of less than a couple of hundred million data points iDEM should perform adequately on such smaller systems. If larger areas or higher density information is required disk space becomes the primary constraint. Processing power is a secondary constraint if time requirements are not restrictive. A smaller system with additional hard drive capacity in the multi-terabyte range extends the capacity of iDEM. Ideally a server system with extensive hard disk capacity or expandable capacity is desirable.

A disk space analysis has been performed to get a general idea of the eventual data storage capacity required to operate a mature version of the system being developed. The space requirements of the mature system are in the order of $\frac{1}{3}$ and $\frac{1}{2}$ a petabyte when iDEM is:

- fully functional,
- loaded with all the input data required for tsunami modelling along the BC coast,

- storing all the intermediate data used to create DEM with DCM,
- storing all the DEM required for the tsunami modelling and,
- has as much free disc space as the database occupies to use as working space

This range of the space requirement encompasses a variety of assumptions as to the number of basic optional functions and level of enhancement incorporated into the library. Details of the analysis can be found in appendix A.2 Disk Space Analysis.

Chapter 5

Evaluation

Evaluation of iDEM's design requires a qualitative approach at this time. A quantitative analysis can be performed once the critical path of the system is operational. At that time qualitative analysis can assess ease of use along with user acceptance of the promoted practices and standards. Quantitative analysis can evaluate processing speed, storage requirements based on addition use data, and output comparison with expectations and existing applications. The following table is a list of the goals as defined by the system requirements, actions taken during the design, and expected results of the decisions driving those actions. iDEM is fully expected to meet the requirements.

Specific reference to applications and functions such as CUBE and ad hoc merging algorithms reflect the initial selection for implementing the iDEM prototype. Later versions allow for enhancement or replacement of these modules.

goal	action	result
modularization		
flexible system framework	a) reconfigurable execution calls by editing ini file b) define distinct functions as external modules	allows same framework to be adapted to various purposes
leverage existing applications	wrap applications to access desired algorithms as separate sub-modules	allows transparent access to algorithms embedded in external applications
implement outstanding required algorithms	code algorithms not found in suitable applications as separate sub-modules	allows wrapped applications and native sub-modules to be treated as equivalent entities
flexible algorithm integration	a) treat each algorithm required to implement a functional module as a separate algorithm sub-module b) replace, upgrade or enhance an algorithm by editing, or adding new,sub-modules	a) allows any particular integrated application to be replaced by a functionally equivalent one with minor re-coding b) allows any particular algorithm to be replaced by an equivalent one with minor re-coding
automation	eliminate expert training	a) eliminate most cleaning b) predefined process c) output quality largely dependent on input d) higher data distortion
	populate uncertainty	a) enable accuracy tracking b) enable data sorting
	ad hoc fusion process	new algorithm a) discard lowest accuracy b) intelligent merging c) user controls acceptable resolution distortion

goal	action	result
stores all input data cleaned points existing DEM raw points	relation database	a) inbuilt GIS functions b) enforce common - coordinate system - position datum
	ensure common format	raster translated to vector
	library of input data	a) avoid input duplication b) enable recreation
creates and stores DEM of desired location at desired resolution	best practice a) grid with cleaning b) grid and interpolate	complimentary algorithms a)CUBE b)Modified Three Step
	enforce tile positioning default grid sizes	enable point position interlock a) promote point interlock
	  library of DEM	b) promote standardization c) enable tessellation a) avoid reprocessing
		b) promote standardization of input for modellers c) enable tessellation and cropping process to define new custom DEM d) enable DEM enhancement merge with new data directly, not reprocess all input data again
DCM for each DEM	propagate uncertainty	a) highlight DEM attributes b) promote trust
complete metadata input processing output	track all data events	a) promote trust b) enable recreation c) product history d) track primary product destination / user
minimize development costs time-lines maintenance	a) integrate existing applications into a new system b) favour OpenSource	a) license costs minimized b) coding need minimized c) time-lines reduced d) low resource requirement e) leverage crown source component maintenance

5.1 Future Considerations

5.1.1 Prototype and Version 1

At this point iDEM's design overview and process flowcharts are fairly detailed. Implementation of a prototype of iDEM's Stage 1 - Data Homogenization is functional though not complete. Three hundred million plus data points have been imported and homogenized. The existing prototype was used to test design ideas, as a result Stage 1 development lags behind the design process.

iDEM is an ongoing project that will migrate to a server environment with flexible configuration. This will allow easier adaptation to handle changing hardware requirements. Once the migration has taken place work will proceed to prototype Stage 2 and Stage 3 of the system.

Integration of applications into the iDEM framework will require flexible interface definitions. These definitions will be the connections between the python wrapped external applications and the controlling python framework. The definitions will create modularization allowing any python wrapped external application, with the correct functionality, to be utilized to fulfill a particular task. This will add flexibility to future upgrades, enhancements and maintenance by allowing most tasks within the process to be isolated and altered without effecting the rest of the system.

In order to achieve a functional system in a timely manner the design process will continue to add enhancements throughout the implementation process. The resulting prototype is expected to function but will not be an implementation of the final design. Once the prototype is functional work should commence on a full version implementing the actual design. This work will, in effect, be an enhancement of the existing code to create a newer version. While the enhancement work is ongoing the functional prototype can be used to produce DEM.

5.1.1.1 Uncertainty Estimation

An application that calculates grid point uncertainty should either be located or developed. If the generalized CUBE uncertainty method covered by patent [38] has an OpenSource implementation or is available under licence this option should seriously be pursued. The method offers increased level of trust as well as processing efficiency.

5.1.1.2 Tidal Datum

Access to existing tidal datum grids of the primary area of interest (BC coast) needs to be obtained to enable the implementation of Stage 2 and Stage 3. A significant facet of existing tidal grids that requires investigation is whether their elevation adjustment figures have associated uncertainty values.

5.1.2 Version 2 and Beyond

5.1.2.1 Outlier Preprocessing - Cleaning Data

Preprocessing of data can encompass many goals. Algorithms with the intention of removing outliers can be implemented in the data homogenization module. As with the proposed uncertainty estimation other processes can examine the datasets looking for outliers. Rather than deleting perceived errors the data should be flagged for user inspection. Flags can be used by an operator to make decisions about including the suspect data in further processing. Such outlier detection efforts are not part of this phase of the project; however provisions should be made for such future expansion.

5.1.2.2 Tidal Inference

The prototype and first full version of iDEM will rely on existing tidal model adjustment grids. In the future, iDEM will implement estimation methods to infer approximate tidal grids or models for regions where they ajoin existing tidal adjustment grids but the tidal model is lacking.

5.1.2.3 High Density Topographic Data Processing

The current design does not include an option to directly process high density topographic data. At present such data is usually LiDAR in nature. The current design assumes that high density topography data is preprocessed into lower density DEM before being ingested by iDEM.

The addition of LiDAR and other high density processing options would further extend iDEM automation and applicable data sources. It would also increase the general applicability of iDEM as a 'one stop DEM option'.

5.1.2.4 Analysis and Comparisons

Functions to analyse DEM statistically, mathematically and visually could assist in research and future iDEM enhancements. Such functions are already implemented in other applications. Some thought should be given to whether iDEM might benefit from integrating additional functions or if they should be used as standalone applications.

Tools to compare different DEM of the same area would be an asset. This would allow the effect of different datasets and/or processing options to be directly investigated.

Comparing tools would allow investigation of how DEM resolution or grid size effects uncertainty estimates of the individual elevation points.

5.1.3 User Friendly Interface

The iDEM system is a `python` library requiring access to the applications integrated into its process. When properly registered with the operating system iDEM functions can be executed from the command line.

A friendly GUI can add a guiding process to assist users in operating iDEM. Such a process could make the various initialization files transparent, reduce training requirements and provide situational aware help functionality.

5.1.4 Data Security

A GUI could also enhance security by restricting direct access to iDEM commands and data since it would require users to have accounts that limit access as desired. Such an enhancement could allow the development of a sandbox environment with various different levels of access to iDEM functions, DEM, DCM and metadata.

5.1.5 Beyond Spherical Modelling

The use case of DEM customized for tsunami modelling has been the major objective behind iDEM. This example has shown iDEM configured with a spherical default coordinate system and accompanying datum. The algorithms selected and applications integrated into the system are configured to handle global spherical data. The application of some minor adjustments to these default conditions and the additional integration of a non-bathymetric high density data gridding subsystem could just as effectively be used to create customized planar DEM. This enhancement extends the use of the iDEM concept to modelling planar structures.

5.1.6 Staying on Course

Careful consideration should be given to any proposed enhancement. Will the proposed enhancement increase iDEM accuracy without decreasing automation? Will it increase automation?

iDEM is conceived to be lean and automated. The most user intensive task is populating the database. The average user should only need to know how to export an existing DEM or how to define a tile to automatically create an new DEM.

Chapter 6

Conclusions

DEM creation can be broken down into two major tasks: data cleaning and data gridding. This proposal for the iDEM framework makes it possible to automate the gridding task. The iDEM design allows non-experts to create customized digital elevation models (DEM) complete with digital certainty model (DCM) and metadata history. The system is designed to only perform rudimentary automatic data cleaning making the output quality largely dependent on the accuracy of the input data. The potential for inaccuracy is reflected in the accompanying DCM, allowing targeted improvement in input data accuracy through judicious preprocessing (outside the iDEM system). The preprocessed and improved dataset can then be reloaded and a more accurate DEM automatically created.

The prototype design of the iDEM framework was conceived as a solution to creating DEM for tsunami modelling. Details of the trade-offs in that domain have been included in this work. The framework would be applicable to the creation of customized DEM of any surface spanning scale from macro to micro.

The automation of customized DEM is accomplished by integrating functions from existing applications as subsystems to build a new system designed to perform a single task. The new iDEM controlling framework ties the integrated applications together into a cohesive python library, along with a new data merging algorithm. The simple algorithm makes it possible to control the acceptable level of dissimilar resolution distortion while automating the merging process.

In the tsunami modelling example, we show how input data can be transformed and stored with a standard coordinate system, position and elevation datum. To minimize uncertainty and maximize flexibility, tidal datum are not standardized prior to storage but they are standardized prior to merging and DEM creation. All default

DEM can be standardized to mean sea level when created and stored in iDEM. When a DEM is exported, it can then be automatically adjusted to the desired tidal datum. This demonstrates the kind of flexibility in design that could be desirable in other use cases as well. The creation of DEM in this example could use an amalgamation of three existing fusion methods that would allow tiling without edge distortion and propagate data uncertainty to create a DCM for every DEM.

A complete mature version of iDEM will be storage intensive. Input data, DEM, DCM, and metadata for the tsunami modelling example covering the west coast of British Columbia (Canada) is projected to require between a $\frac{1}{3}$ and $\frac{1}{2}$ a petabyte depending on which highest resolution is utilized for DEM.

The challenge of needing data in different formats to be easily integrated into DEM with certainty was exemplified in the tsunami example. The design strategy of iDEM is to automatically generate customized DEM based on the desired data fusion techniques and optimal measurements available. However, reconciling the differences between formats will still require customized attention on an application-specific basis.

Appendix A

Additional Information

A.1 Glossary

3Step	Three Step Approach.
ANSI	American National Standards Institute.
API	Application Program Interface, is a set of routines, protocols, and tools for building software applications.
artifact	Something observed in a scientific investigation or experiment that is not naturally present but occurs as a result of the preparative or investigative procedure.
ASTER	Advanced Spaceborne Thermal Emission and Reflection Radiometer.
autocorrelation	Correlation between the elements of a series and others from the same series separated from them by a given interval.
AUV	Autonomous Underwater Vehicle.
BAG	Bathymetric Attributed Grid, a raster data file format capable of holding meta-data, elevation values and uncertainty.
BC	British Columbia.

CHS	Canadian Hydrographic Service.
Constrained	[pertaining to gridding] Algorithms that are unable to predict values outside the range of the sample data.
correlation	A quantity measuring the extent of interdependence of variable quantities.
CSV	Comma Separated Values.
CUBE	Combined Uncertainty and Bathymetric Estimator.
datum	a fixed starting point, a reference.
DCM	Digital Certainty Model, where certainty refers to a level of trust that is conveyed by the uncertainty values presented by the model. The DCM values represent the accuracy of the DEM for which the DCM is an uncertainty model.
DEM	Digital Elevation Model.
ellipsoid	A three-dimensional figure whose plane sections are ellipses or circles.
Free	[pertaining to gridding] Algorithms that can interpolate values beyond the range of the sample data.
GDAL	Geospatial Data Abstraction Library.
GEBCO	General Bathymetric Chart of the Oceans.
geoid	The hypothetical shape of the earth, coinciding with mean sea level and its imagined extension under (or over) land areas.
geocentric	Measured from or considered in relation to the center of the earth.
GIS	Geographic Information System.
GMT	Generic Mapping Tools.
GPS	Global Positioning System.
GUI	Graphical User Interface.
iDEM	integrator of Digital Elevation Models.

IHO	International Hydrographic Organization.
Inverse Distance to a Power	Another name for the Inverse Distance Weighted (IDW) gridding method.
ISO	International Organization for Standardization.
isotropic	Having a property that has the same value when measured in different directions.
Kriging	An interpolation technique that also provides variance, named for D. G. Krige.
liboncbf	library Ocean Networks Canada bathymetric fusion.
LiDAR	Light Detection And Ranging.
Locally Weighted Polynomial Regression	Another name for Local Weighted Smoothing Regression.
M3Step	Modified Three Step Approach.
MBES	MultiBeam Echo Sounder.
Minimum Curvature	Another name for the Spline Interpolation.
NOAA	National Oceanic and Atmospheric Administration (USA).
ODBC	Open DataBase Connectivity.
OGR	related to GDAL - OGR Simple Features Library.
Olex	Norwegian system for navigation, fishery plotting and ocean mapping.
ONC	Ocean Networks Canada.
oncbf	Ocean Networks Canada bathymetric fusion.
orthometric	A height measured with reference to the geoid.

outlier	In statistics, an outlier is an observation point that is distant from other observations. An outlier may be due to variability in the measurement or it may indicate measurement error; some such errors can be detected and excluded from the data set.
predict	The calculation of a value at a particular spacial location based on existing sample values within a sample area.
prediction surface	A surface defined by a collection of predicted values.
ROV	Remotely Operated Vehicle.
SAGA	System for Automated Geoscientific Analyses.
sample	A measured value.
sample area	The spacial area (2D or 3D) in which samples are taken.
SBES	Single Beam Echo Sounder.
SRTM	Shuttle Radar Topography Mission.
sandbox	An isolated computing environment: • In general: used by software developers to test new programming code. • In relation to this research: used by model developers to test and compare new models utilizing the same DEM.
SQL	Structured Query Language.
TPU	Total Propagated Uncertainty.
USGS	U.S. Geological Survey.
UTM	Universal Transverse Mercator coordinate system.
VBES	Vertical Beam Echo Sounder.

wrt	short for *with respect to*.
Quadratic Least Squares Regression	Another name for Local Weighted Smoothing Regression.
zone	[pertaining to gridding] A smaller defined area contained within a sample area.
zone	[pertaining to UTM] An area in which a particular set of UTM projection parameters are valid.

A.2 Disk Space Analysis

A.2.1 Estimating the number of data points

The primary purpose driving iDEM development is the creation of DEM for tsunami modelling along the west coast of British Columbia (BC). This means that the eventual scope of the project encompasses a large area of the Earth's surface. In order to be useful as an input for tsunami modelling the DEM need to include places on the planet that can be sources of tsunamis which may impact the west coast of BC and all the waters between those sources and that coast.

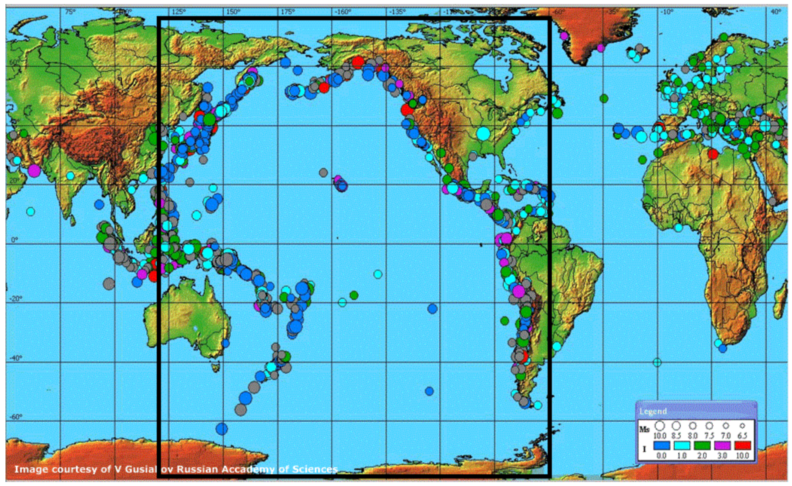


Figure A.1: SRTM30 data covering the Pacific Ring of Fire [53]

For the purpose of projecting and analysing storage requirements for a well populated instance of the system the required elevation information can be considered as two distinct datasets:

- Low resolution data of possible tsunami locations and the waters between them and the coastal area of BC. For BC the main tsunami sources relate to the Pacific Ring of Fire and can be far-field, or near-field.

- Higher resolution information of the waters along the coastal region of BC. This includes all the waters from the edge of the continental shelf including the shoreline and the topography of the coastal region.

The initial dataset chosen for the low resolution region is SRTM30 v10. This is a 30 arc-second regular grid containing both bathymetric and topographic data. This information is available as one global dataset or as smaller subsets each covering a specific area of the globe. In order to reduce the required storage space it is envisioned that only the subsets relevant to the BC coast will be utilized. Figure A.1 shows an approximate outline of the area of interest. This area does not necessarily encompass only the required subsets of SRTM30 but it is a starting point for an approximation of storage requirements.

To be able to model near and far-field for the BC coast the area of interest is framed by the corner coordinates $[(120, 90)(-60, 90)\ (-60, -90)(120, -90)]$ and covers half the globe; 180° longitude 180° latitude (see Figure A.1). Since SRTM30 has a 30 arc-second resolution there are 120 x 120 grid cells per square degree. The area of interest contains $0.47G$ points of SRTM30 data.

Figure A.2: Coastal region of British Columbia [42].

The higher resolution area of interest is outlined in Figure A.2. It is framed by the corner coordinates $[(-135, 55)(-128, 55)(-120, 47)(-126, 47)]$ and covers 52 square degrees. For the coastal region DEMs will be required at various resolutions. Commonly utilized resolutions measured in arc-seconds are $9, 3, 1, \frac{1}{3}$ and $\frac{1}{9}$. Table A.1 shows the resolutions, and the number of points required to obtain that resolution per square degree. The assumption is that a fully mature database will contain DEM covering the whole area of interest at all resolution.

resolution			points
degrees	X meters*	Y meters*	per deg^2
9	179.2	278.1	0.16M
3	59.7	92.7	1.44M
1	19.9	30.9	12.7M
1/3	6.6	10.3	0.12G
1/9	2.2	3.0	1.05G

* : @ 128° longitude & 50° latitude [113]

Table A.1: Common DEM resolutions.

For this analysis the following assumptions have been made about the coastal region:

- Input data (raw elevation data & existing DEM) have full coverage at all common resolutions.
- Output DEM will have full coverage of the coastal region at all common resolutions.

These assumptions mean that the coastal region will require 61 Giga-points of input data and generate another 61 Giga-points of output.

The following assumptions were made about what would be stored:

- An average of three different methods will be used to estimate uncertainty of each point.
- Each point will have uncertainty estimates calculated at each grid resolution.
- Complete copies of tile definition input data utilized to create each output DEM will be stored to ensure complete duplication of prior work.
- An average of three DEM will be created form each area at each resolution.
- Each DEM created will be stored.
- Metadata will be negligible compared with the input and output data.
- The first generation of the system will assume that all input tide levels are homogeneous and that output will be fixed to the same datum; i.e. no tidal data will need to be stored.

A simplified estimate of the space requirement based on this information can be seen in table A.2.
If the database is enhanced with data and DEM of $\frac{1}{27}$ of an arc-second (0.7 meters) in shallow waters near the coast and sensitive coastal land areas; and if an estimated 20% of BC coastal region will be represented at this resolution then the required space increases by 14TB.

These simple analysis options are summarized in table A.3:

Data description	TB
Input data	5
Uncertainty estimation	28
Tile definitions	14
Tidal data	0
Output DEM	14
Total	61

Table A.2: Simplified estimates of data storage requirements.

Data description	TB
Basic system	61
Enhanced	75

Table A.3: Simplified estimates of data storage options.

A.2.2 Transforming simple space estimates into practical numbers

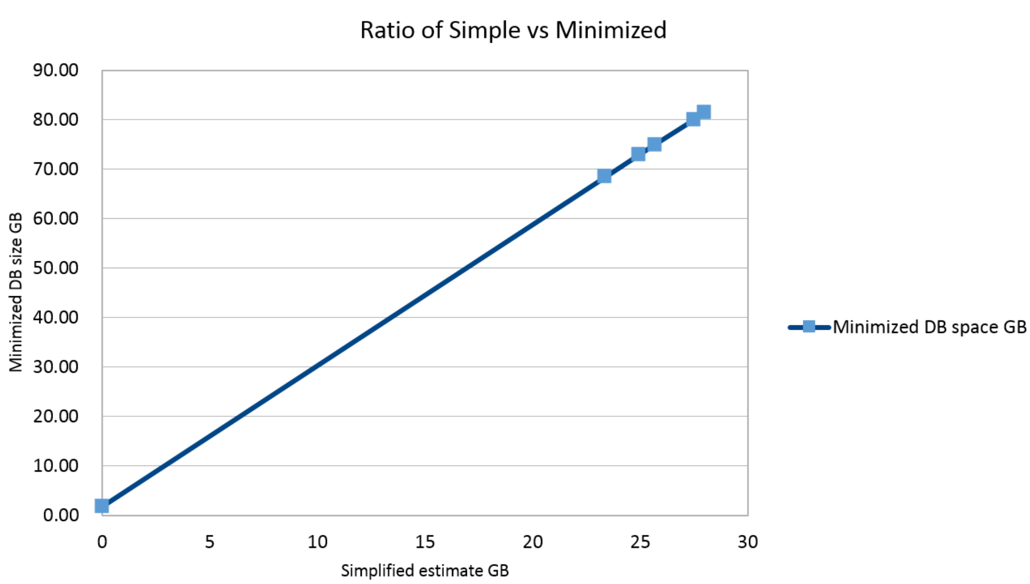


Figure A.3: Simplified estimate versus practical space requirements.

A database requires more space to store information than a simple estimation of bytes in tables would indicate. This is especially true for a database with GIS capability utilizing spacial indexing. Allowance also needs to made for the fact that a quiescent, compacted database (minimized) occupies significantly less disk space than a working one. Figure A.3 shows the results of a comparison of the simple estimation technique and actual disk usage of a PostgreSQL database with PostGIS functionality containing up to 283 million point, no uncertainty, no tile definitions and no tidal information. The comparison reveals that practical requirements are 3 time the simple estimate. In order to observe the disk requirements of the database under working conditions a VACUUM FULL operation was performed. This operation re-indexes and compacts the whole database. Under these conditions the required disk space doubled. This means that a good measure of disk space requirements can be obtained by multiplying the simplified estimates by 6. The results of the practical disk space requirements are shown in table A.4, note the change in units from terabytes to pentabytes.

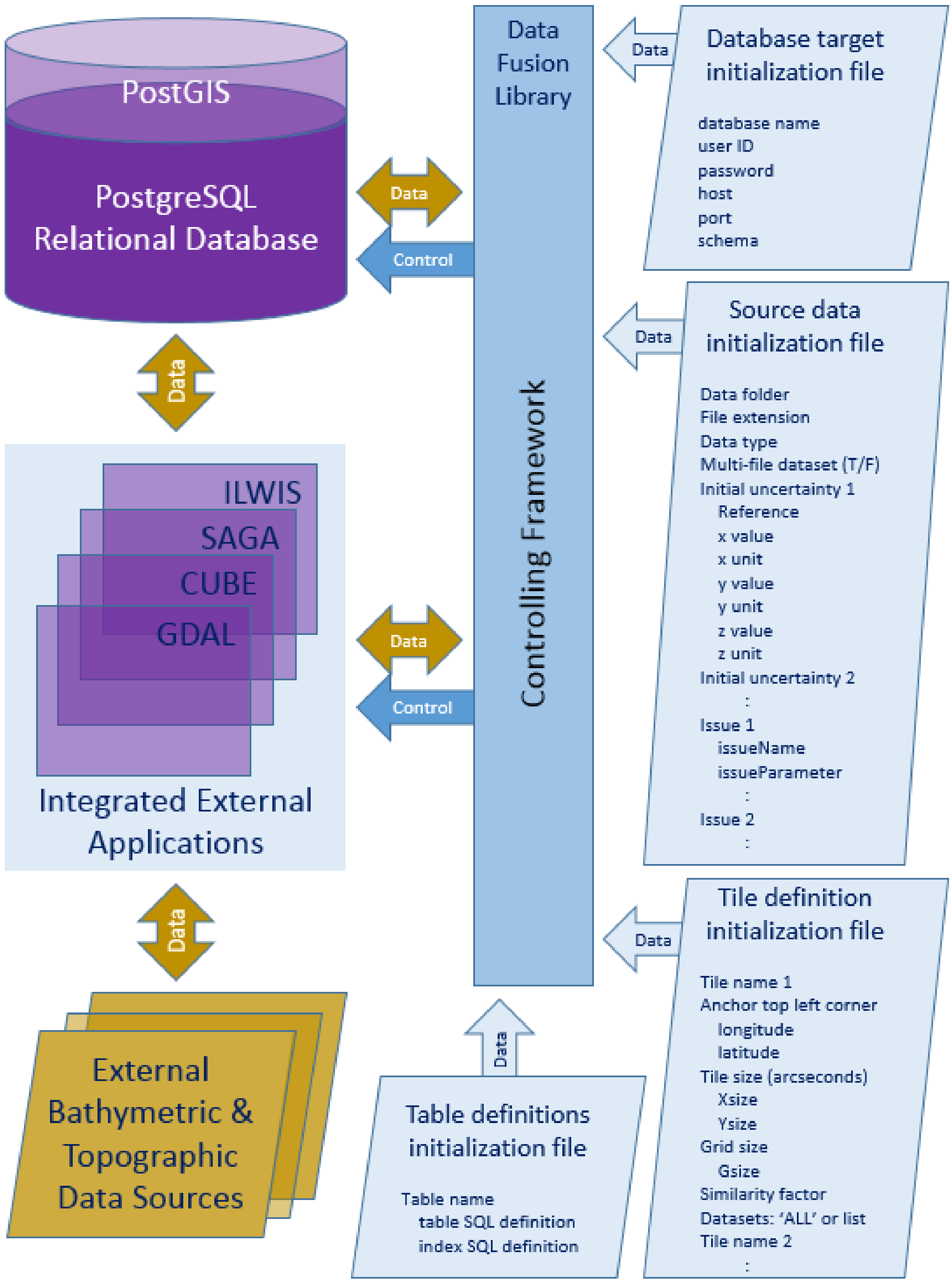
Data description	PB
Basic system	0.37
Enhanced	0.45

Table A.4: Practical estimates of data storage options.

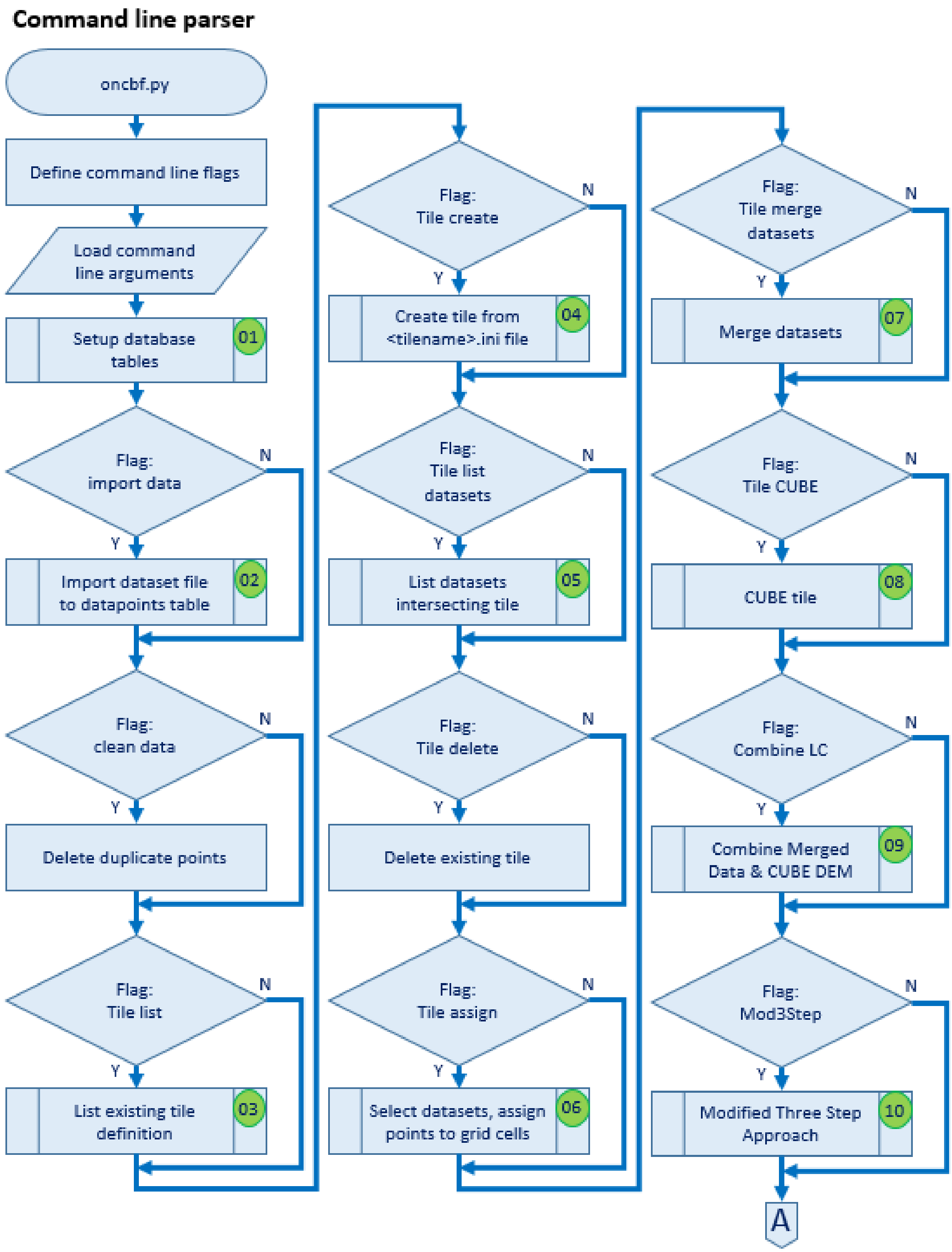
A.3 Design Flowcharts

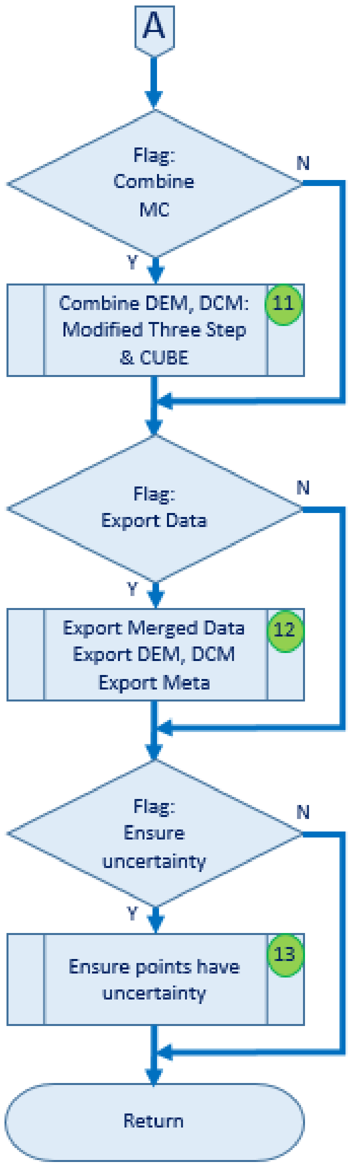
The core of the iDEM is a GIS enabled relational database. A software library with clearly defined functions make up a control framework that integrates and directs the flow, manipulation and tracking of data. The manipulation of the data is directed, and in some cases actually performed, by modules of the controlling framework. Most manipulations attributable to the controlling framework are achieved through SQL interactions with the database. The majority of sophisticated data manipulation is achieved by external applications acting upon the data. Such action is directed by the control framework. Wherever possible the external applications will interface directly with the relational database to manipulate the data under the guidance of the control framework. In the following flowcharts there is no explicit distinction made between the controlling framework and the GIS enabled relational database. The controlling framework interacts with the database directly using SQL commands. Calls to external applications are clearly distinguished from the controlling framework and database through the use of flowchart symbols with a darker purple background. The routines and sub-routines depicted in the flowcharts are referenced by name/title and numbers highlighted in green circles.

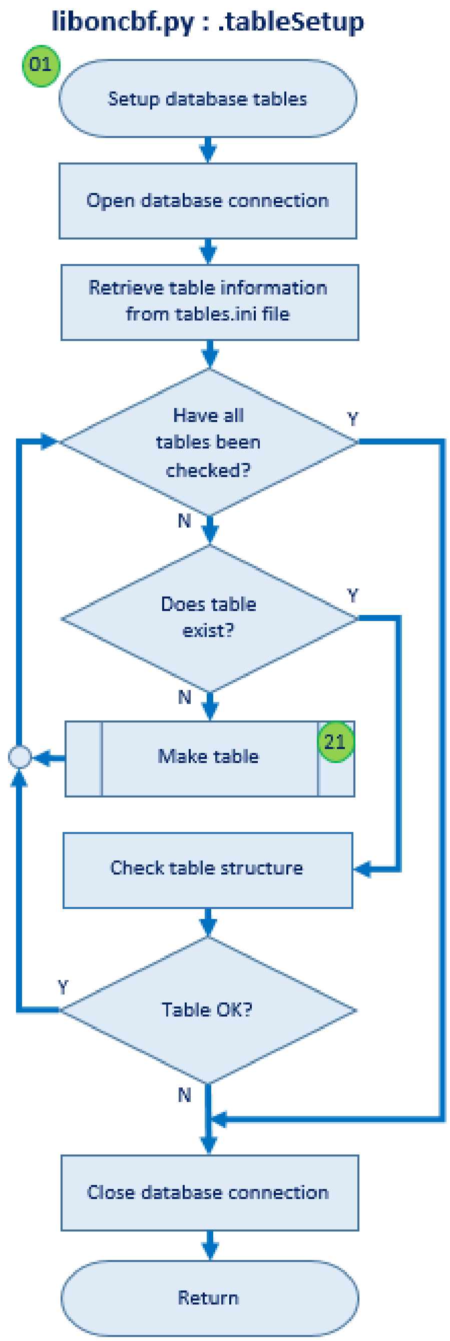
DEM: Digital Elevation Model
DCM: Digital Confidence Model



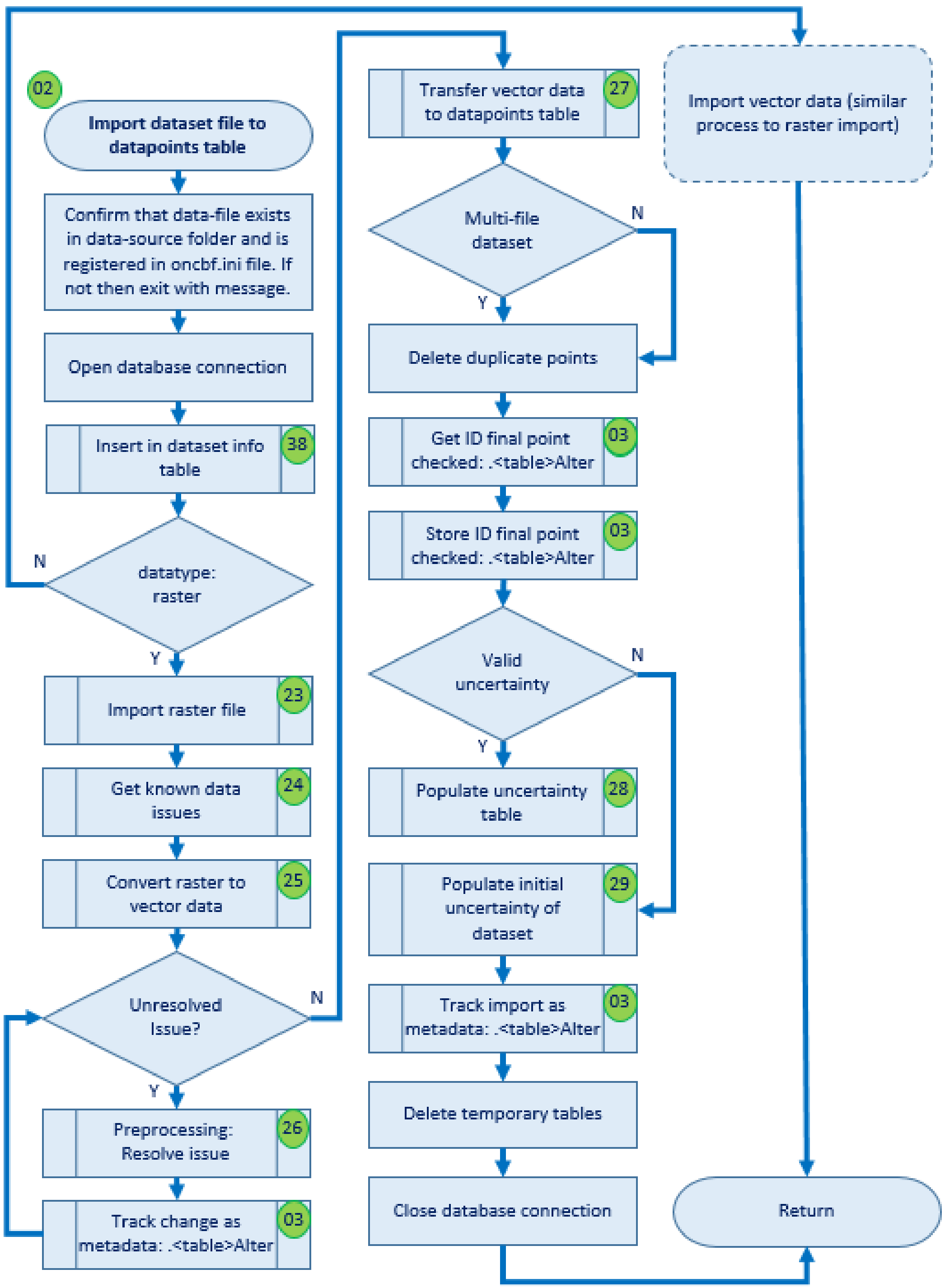
PostGIS
PostgreSQL Relational Database
Data
Control
Data
ILWIS
SAGA
CUBE
GDAL
Integrated External Applications
Data
Control
Data
External Bathymetric & Topographic Data Sources
Data Fusion Library
Controlling Framework
Data
Database target initialization file
database name
user ID
password
host
port
schema
Data
Source data initialization file
Data folder
File extension
Data type
Multi-file dataset (T/F)
Initial uncertainty 1
Reference
x value
x unit
y value
y unit
z value
z unit
Initial uncertainty 2
:
Issue 1
issueName
issueParameter
:
Issue 2
:
Data
Tile definition initialization file
Tile name 1
Anchor top left corner
longitude
latitude
Tile size (arcseconds)
Xsize
Ysize
Grid size
Gsize
Similarity factor
Datasets: 'ALL' or list
Tile name 2
:
Data
Table definitions initialization file
Table name
table SQL definition
index SQL definition

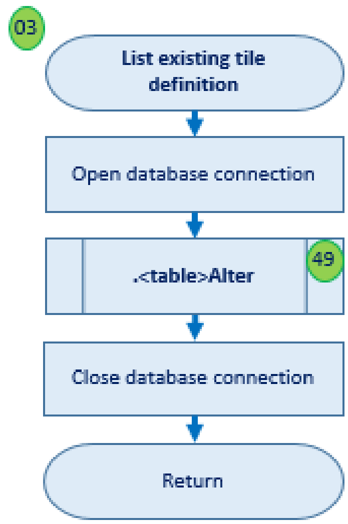

Command line parser
oncbf.py
Define command line flags
Load command line arguments
Setup database tables
01
Flag: import data
Import dataset file to datapoints table
02
Flag: clean data
Delete duplicate points
Flag: Tile list
List existing tile definition
03
Flag: Tile create
Create tile from <tilename>.ini file
04
Flag: Tile list datasets
List datasets intersecting tile
05
Flag: Tile delete
Delete existing tile
Flag: Tile assign
Select datasets, assign points to grid cells
06
Flag: Tile merge datasets
Merge datasets
07
Flag: Tile CUBE
CUBE tile
08
Flag: Combine LC
Combine Merged Data & CUBE DEM
09
Flag: Mod3Step
Modified Three Step Approach
10
Y
N
A

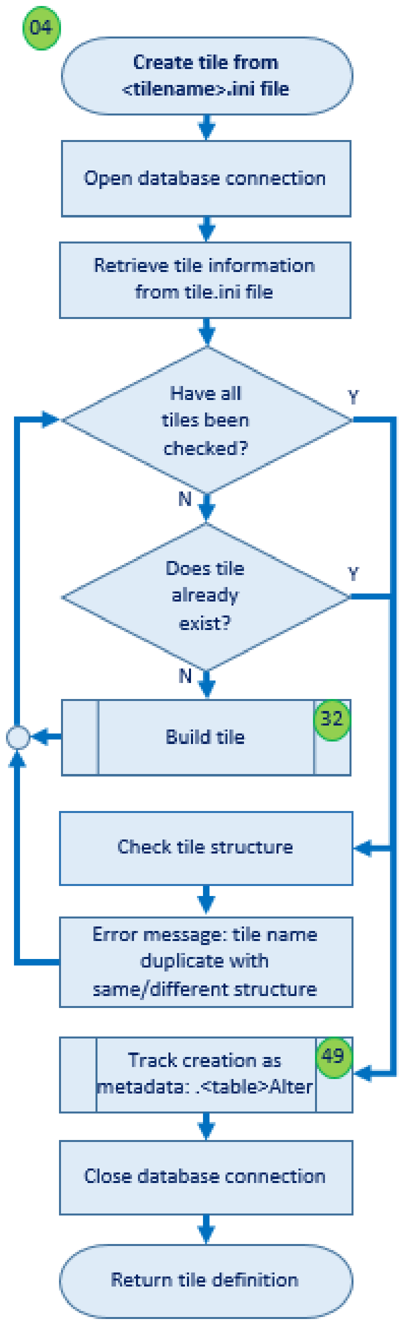

A
Flag:
Combine
MC
N
Y
Combine DEM, DCM:
Modified Three Step
& CUBE
11
Flag:
Export Data
N
Y
Export Merged Data
Export DEM, DCM
Export Meta
12
Flag:
Ensure
uncertainty
N
Y
Ensure points have
uncertainty
13
Return

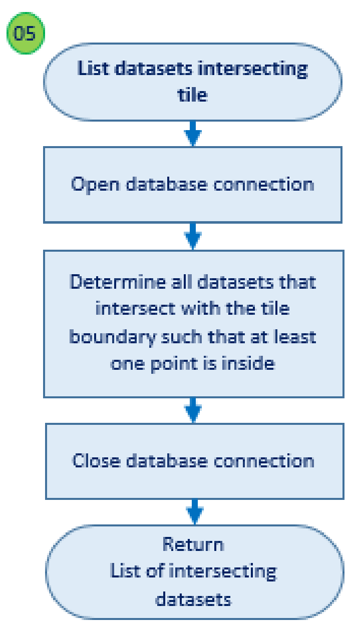

liboncbf.py : .tableSetup
01
Setup database tables
Open database connection
Retrieve table information
from tables.ini file
Have all
tables been
checked?
Y
N
Does table
exist?
Y
N
Make table
21
Check table structure
Table OK?
Y
N
Close database connection
Return

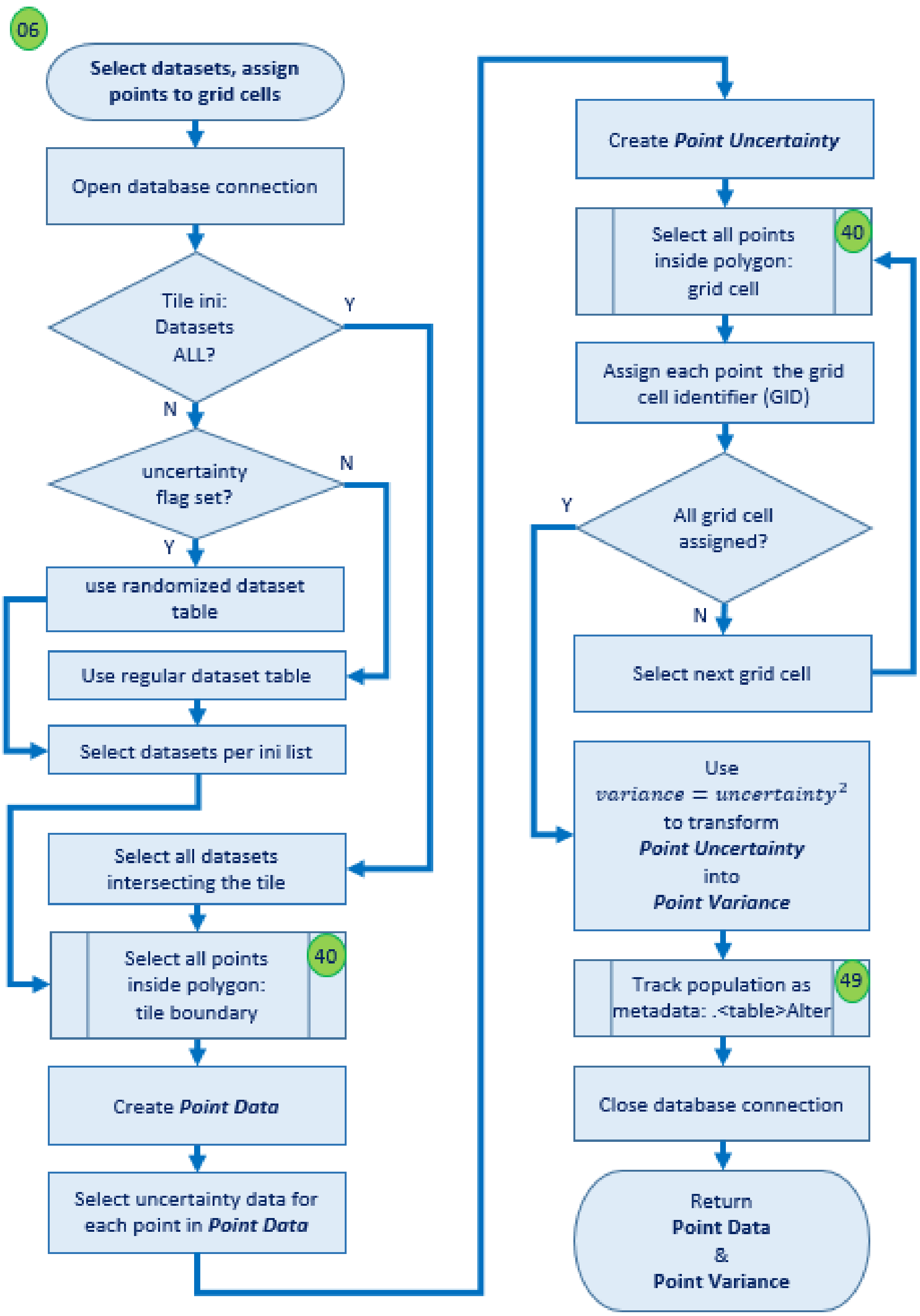
libconcbf.py : .dsFile2datapointsTable

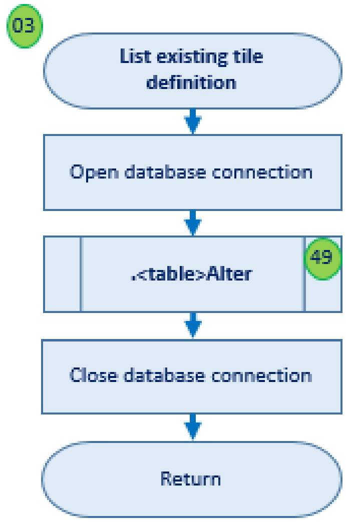


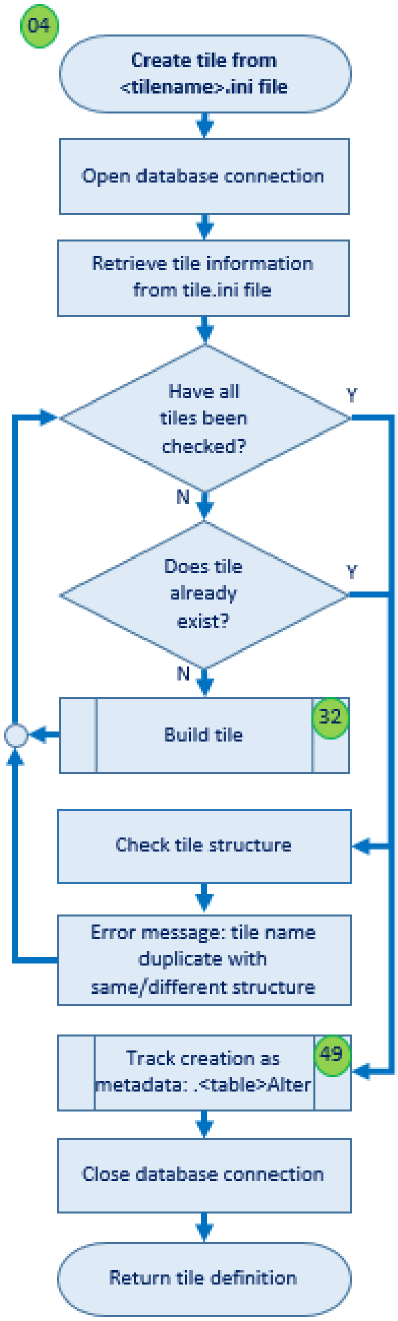

03
List existing tile definition
Open database connection
.<table>Alter
49
Close database connection
Return

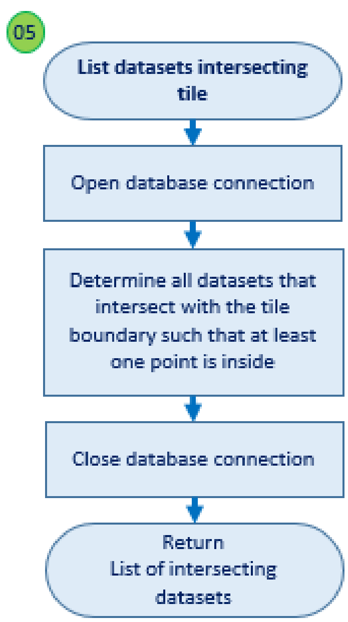

04
Create tile from <tilename>.ini file
Open database connection
Retrieve tile information from tile.ini file
Have all tiles been checked?
Y
N
Does tile already exist?
Y
N
Build tile
32
Check tile structure
Error message: tile name duplicate with same/different structure
Track creation as metadata: .<table>Alter
49
Close database connection
Return tile definition

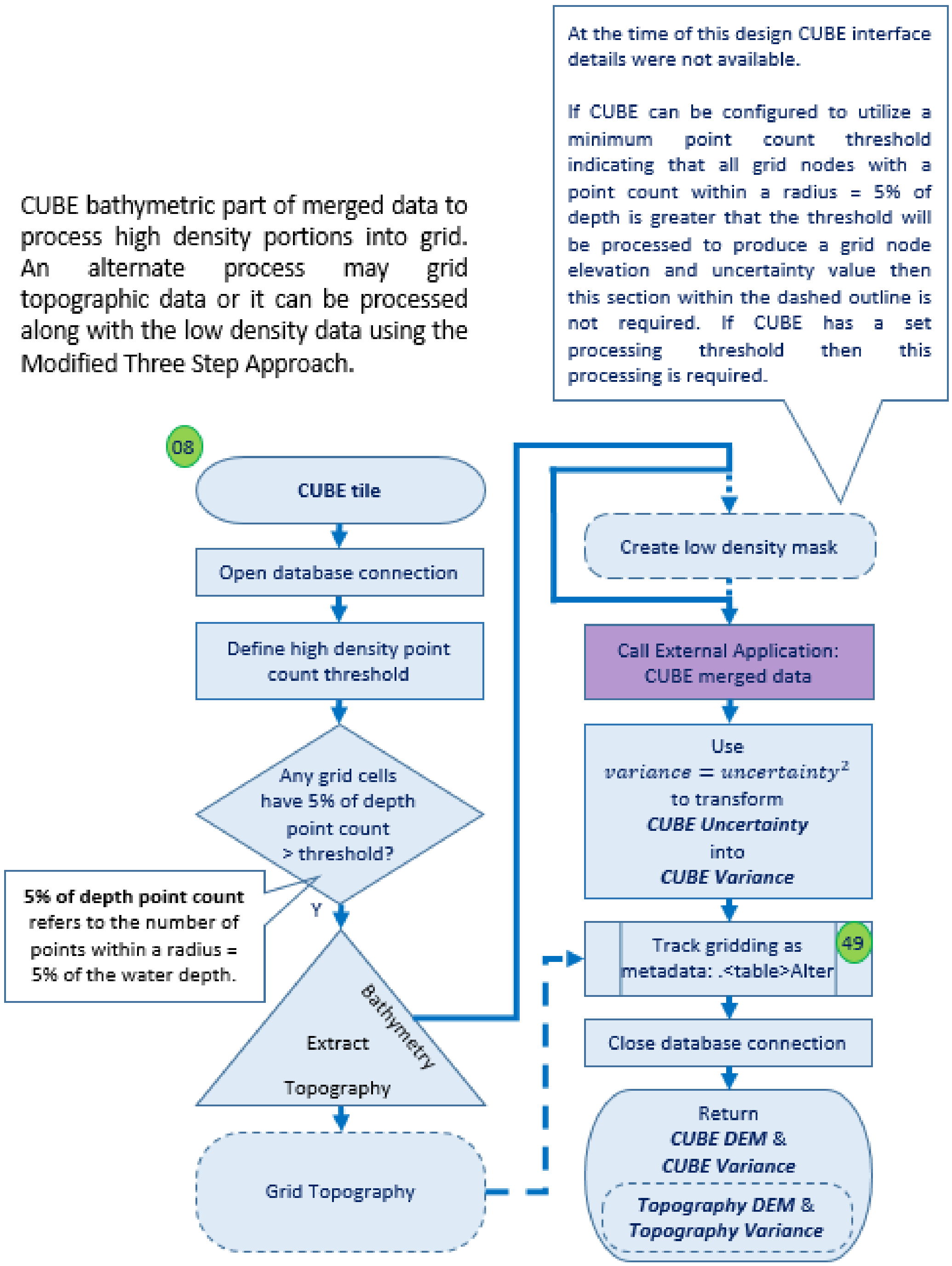

05
List datasets intersecting tile
Open database connection
Determine all datasets that intersect with the tile boundary such that at least one point is inside
Close database connection
Return
List of intersecting datasets

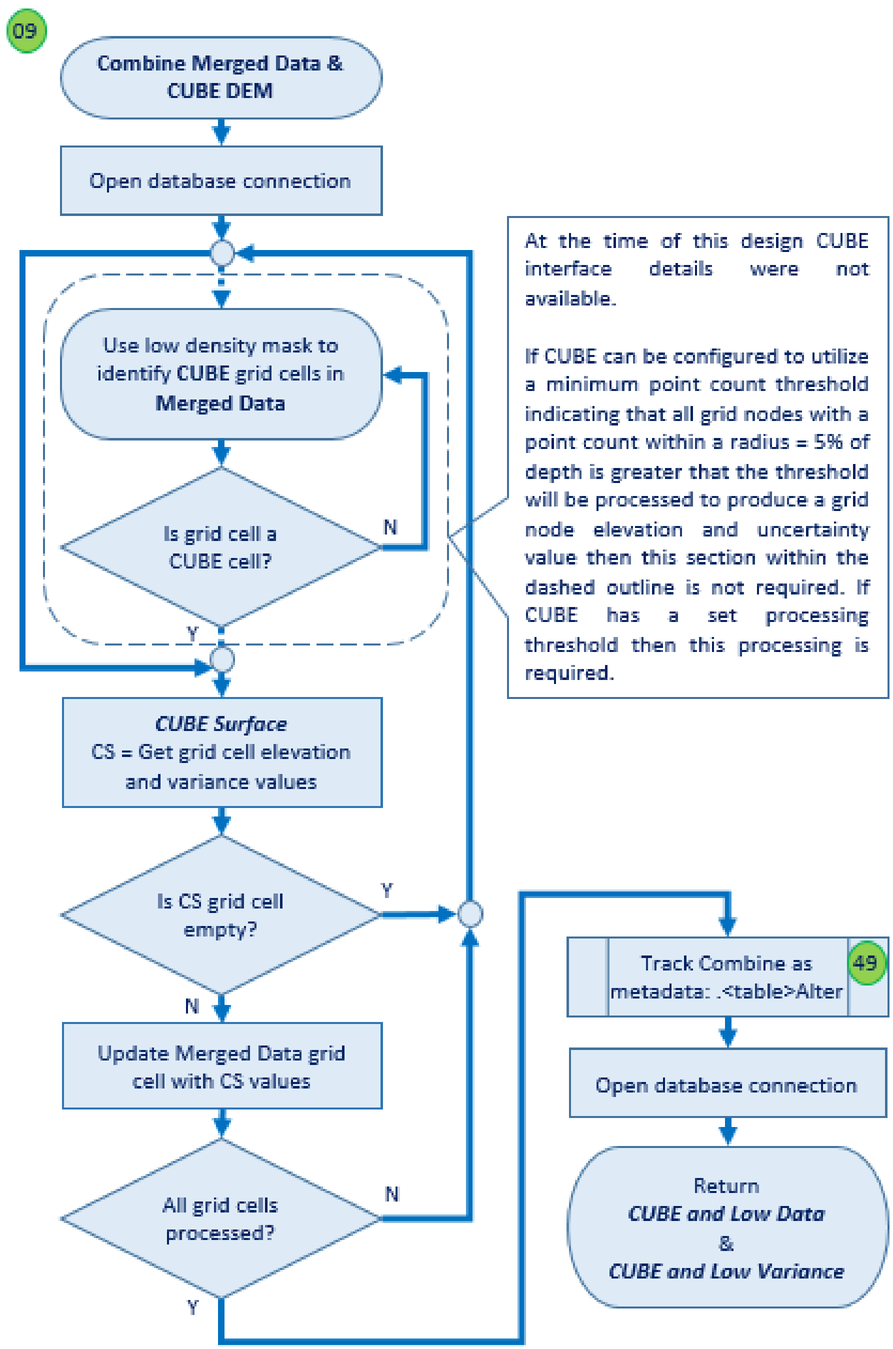

06
Select datasets, assign points to grid cells
Open database connection
Tile ini: Datasets ALL?
Y
N
uncertainty flag set?
N
Y
use randomized dataset table
Use regular dataset table
Select datasets per ini list
Select all datasets intersecting the tile
Select all points inside polygon: tile boundary
40
Create Point Data
Select uncertainty data for each point in Point Data
Create Point Uncertainty
Select all points inside polygon: grid cell
40
Assign each point the grid cell identifier (GID)
All grid cell assigned?
Y
N
Select next grid cell
Use
variance = uncertainty²
to transform
Point Uncertainty
into
Point Variance
Track population as metadata: .<table>Alter
49
Close database connection
Return
Point Data
&
Point Variance

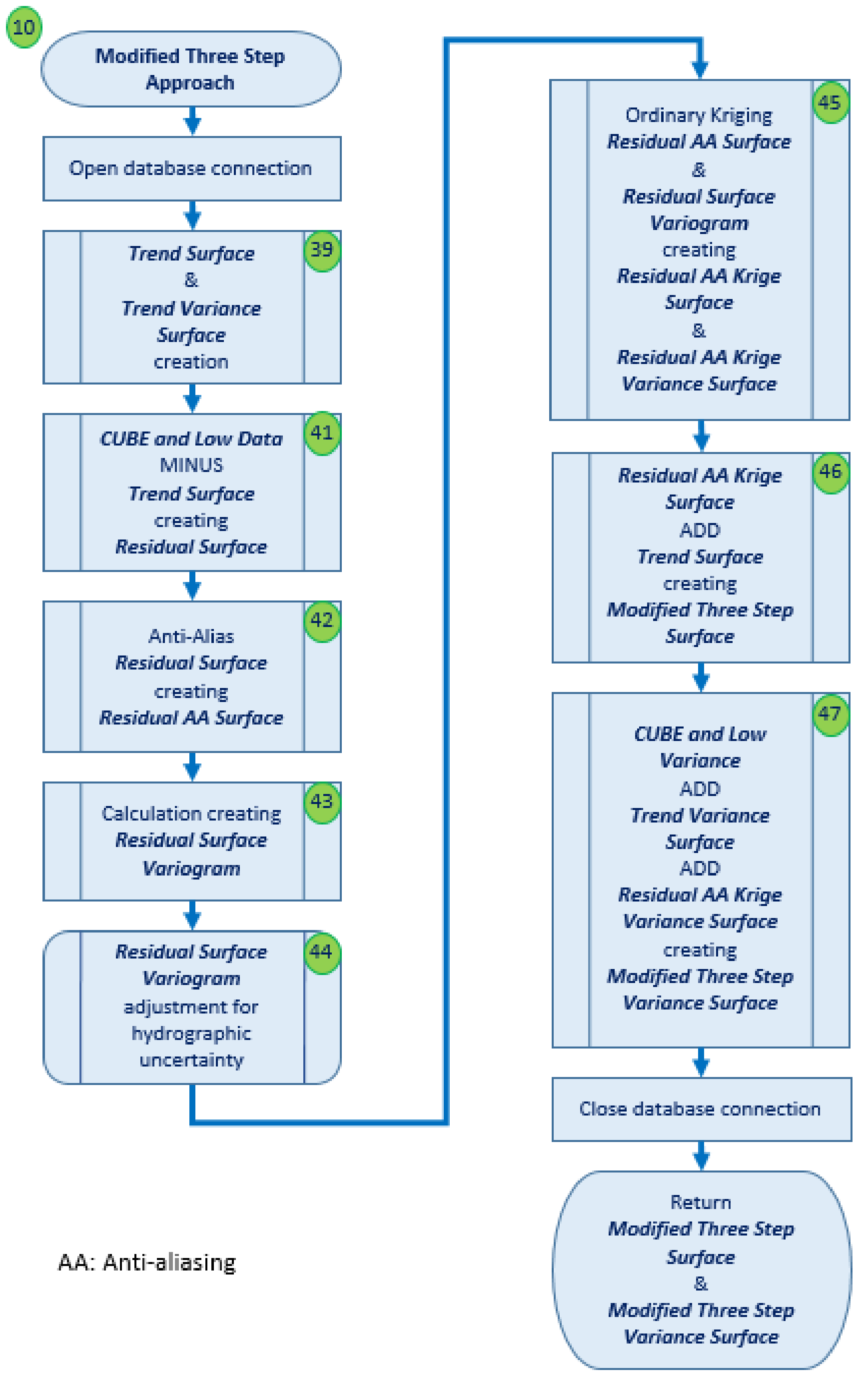

03
List existing tile definition
Open database connection
.<table>Alter
49
Close database connection
Return

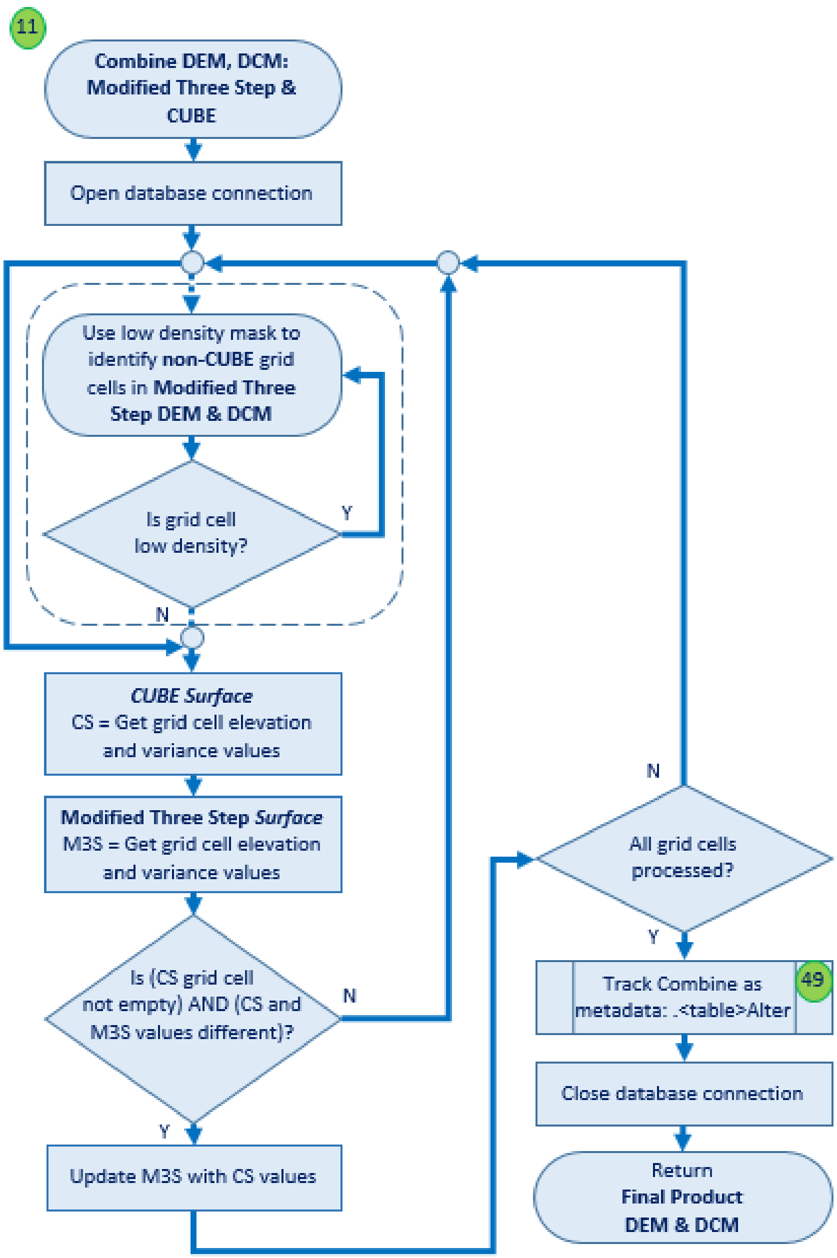

04
Create tile from <tilename>.ini file
Open database connection
Retrieve tile information from tile.ini file
Have all tiles been checked?
Y
N
Does tile already exist?
Y
N
Build tile
32
Check tile structure
Error message: tile name duplicate with same/different structure
Track creation as metadata: .<table>Alter
49
Close database connection
Return tile definition

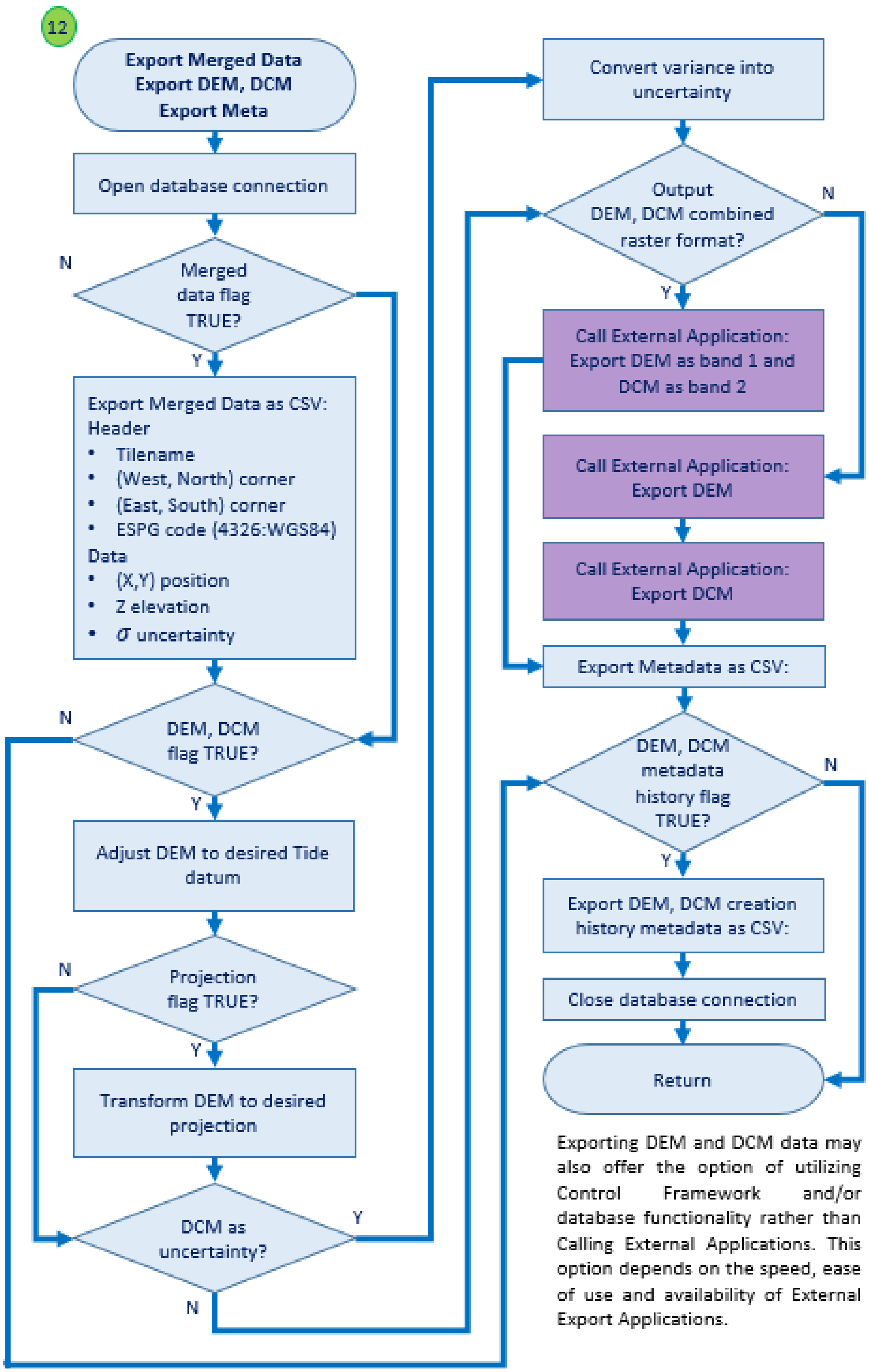

05
List datasets intersecting tile
Open database connection
Determine all datasets that intersect with the tile boundary such that at least one point is inside
Close database connection
Return
List of intersecting datasets

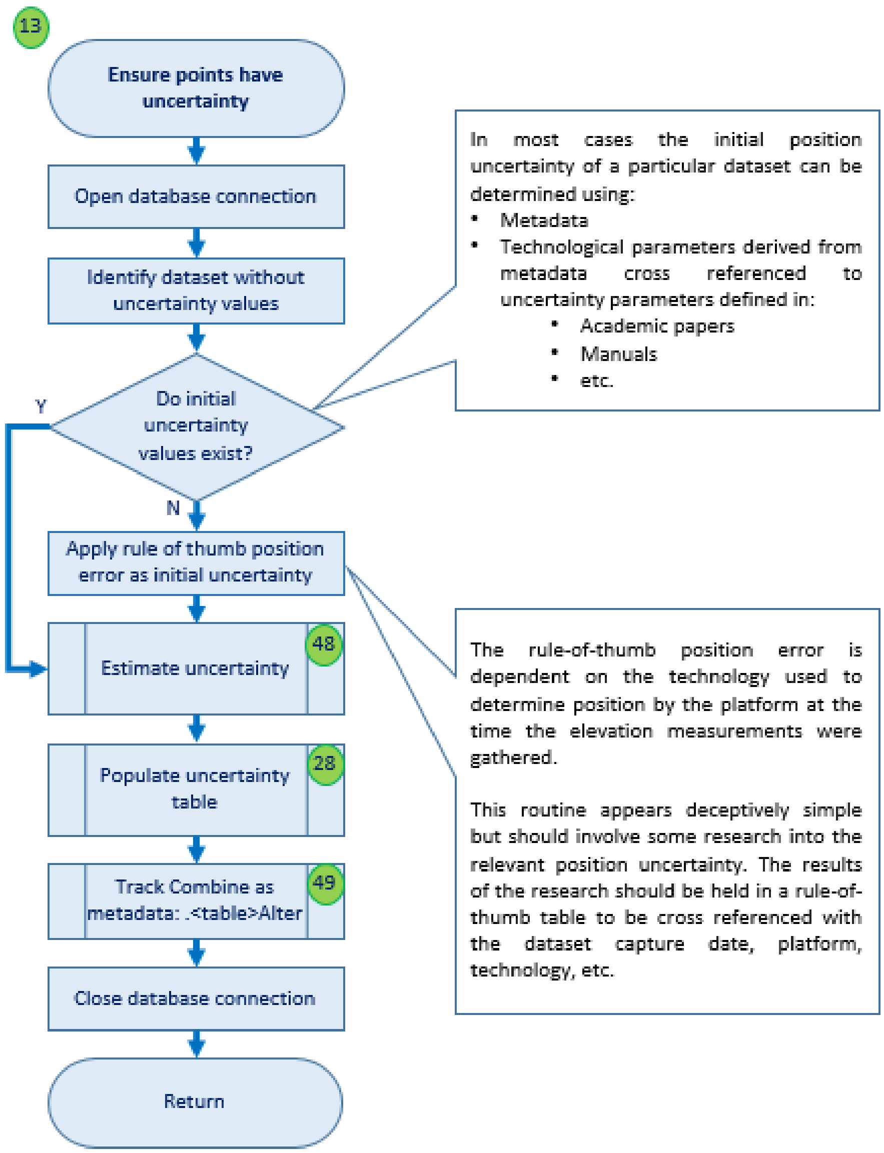

At the time of this design CUBE interface details were not available.
If CUBE can be configured to utilize a minimum point count threshold indicating that all grid nodes with a point count within a radius = 5% of depth is greater that the threshold will be processed to produce a grid node elevation and uncertainty value then this section within the dashed outline is not required. If CUBE has a set processing threshold then this processing is required.
CUBE bathymetric part of merged data to process high density portions into grid. An alternate process may grid topographic data or it can be processed along with the low density data using the Modified Three Step Approach.
08
CUBE tile
Open database connection
Define high density point count threshold
Any grid cells have 5% of depth point count > threshold?
5% of depth point count refers to the number of points within a radius = 5% of the water depth.
Y
Extract
Bathymetry
Topography
Grid Topography
Create low density mask
Call External Application: CUBE merged data
Use
$variance = uncertainty^2$
to transform
CUBE Uncertainty
into
CUBE Variance
Track gridding as metadata: .<table>Alter
49
Close database connection
Return
CUBE DEM &
CUBE Variance
Topography DEM &
Topography Variance

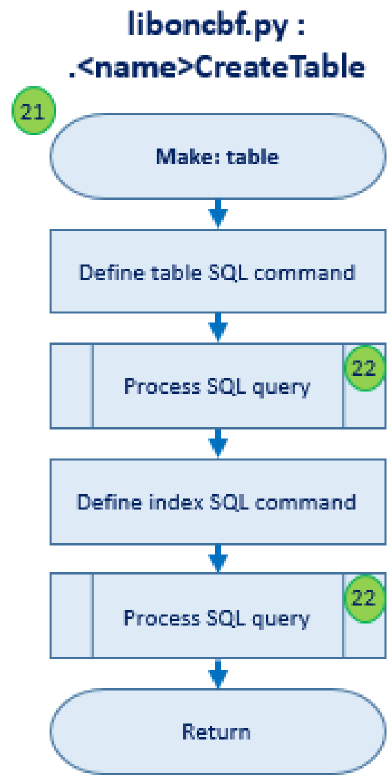

09
Combine Merged Data & CUBE DEM
Open database connection
Use low density mask to identify CUBE grid cells in Merged Data
Is grid cell a CUBE cell?
N
Y
At the time of this design CUBE interface details were not available.
If CUBE can be configured to utilize a minimum point count threshold indicating that all grid nodes with a point count within a radius = 5% of depth is greater that the threshold will be processed to produce a grid node elevation and uncertainty value then this section within the dashed outline is not required. If CUBE has a set processing threshold then this processing is required.
CUBE Surface
CS = Get grid cell elevation and variance values
Is CS grid cell empty?
Y
N
Update Merged Data grid cell with CS values
All grid cells processed?
N
Y
Track Combine as metadata: .<table>Alter
49
Open database connection
Return
CUBE and Low Data
&
CUBE and Low Variance

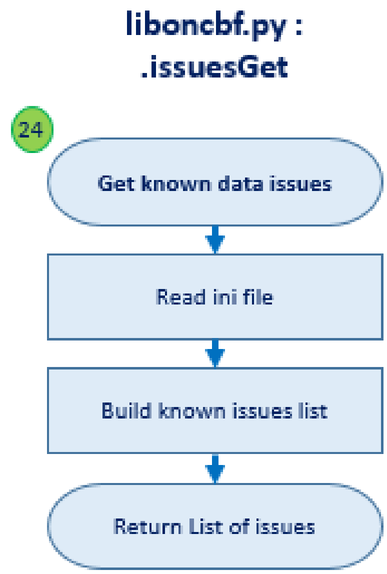

10
Modified Three Step Approach
Open database connection
39
Trend Surface & Trend Variance Surface creation
41
CUBE and Low Data MINUS Trend Surface creating Residual Surface
42
Anti-Alias Residual Surface creating Residual AA Surface
43
Calculation creating Residual Surface Variogram
44
Residual Surface Variogram adjustment for hydrographic uncertainty
45
Ordinary Kriging Residual AA Surface & Residual Surface Variogram creating Residual AA Krige Surface & Residual AA Krige Variance Surface
46
Residual AA Krige Surface ADD Trend Surface creating Modified Three Step Surface
47
CUBE and Low Variance ADD Trend Variance Surface ADD Residual AA Krige Variance Surface creating Modified Three Step Variance Surface
Close database connection
Return Modified Three Step Surface & Modified Three Step Variance Surface
AA: Anti-aliasing

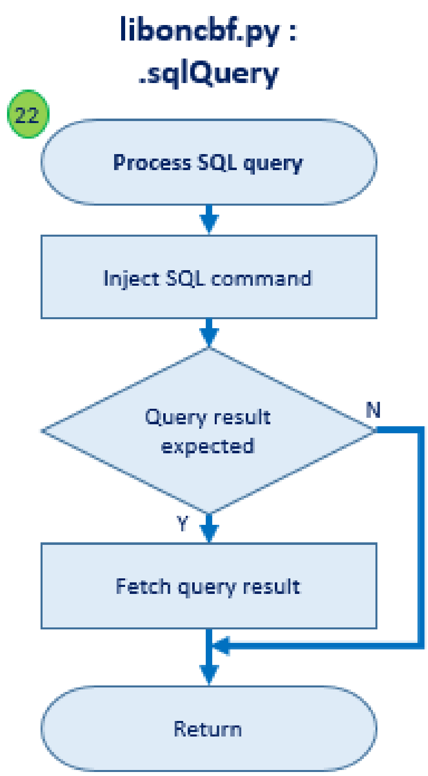

11
Combine DEM, DCM: Modified Three Step & CUBE
Open database connection
Use low density mask to identify non-CUBE grid cells in Modified Three Step DEM & DCM
Is grid cell low density?
Y
N
CUBE Surface
CS = Get grid cell elevation and variance values
Modified Three Step Surface
M3S = Get grid cell elevation and variance values
Is (CS grid cell not empty) AND (CS and M3S values different)?
N
Y
Update M3S with CS values
All grid cells processed?
N
Y
Track Combine as metadata: .<table>Alter
49
Close database connection
Return
Final Product
DEM & DCM


12
Export Merged Data
Export DEM, DCM
Export Meta
Open database connection
Merged data flag TRUE?
N
Y
Export Merged Data as CSV:
Header
• Tilename
• (West, North) corner
• (East, South) corner
• ESPG code (4326:WGS84)
Data
• (X,Y) position
• Z elevation
• σ uncertainty
DEM, DCM flag TRUE?
N
Y
Adjust DEM to desired Tide datum
Projection flag TRUE?
N
Y
Transform DEM to desired projection
DCM as uncertainty?
Y
N
Convert variance into uncertainty
Output DEM, DCM combined raster format?
N
Y
Call External Application: Export DEM as band 1 and DCM as band 2
Call External Application: Export DEM
Call External Application: Export DCM
Export Metadata as CSV:
DEM, DCM metadata history flag TRUE?
N
Y
Export DEM, DCM creation history metadata as CSV:
Close database connection
Return
Exporting DEM and DCM data may also offer the option of utilizing Control Framework and/or database functionality rather than Calling External Applications. This option depends on the speed, ease of use and availability of External Export Applications.

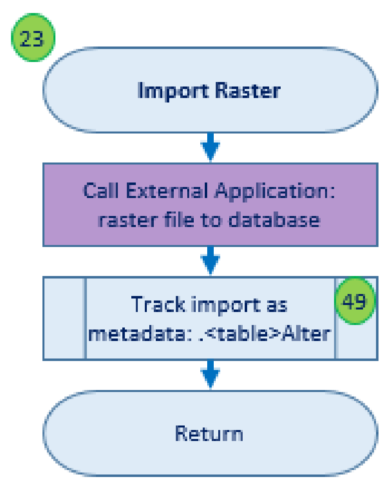

13
Ensure points have uncertainty
Open database connection
Identify dataset without uncertainty values
Do initial uncertainty values exist?
Y
N
Apply rule of thumb position error as initial uncertainty
Estimate uncertainty
48
Populate uncertainty table
28
Track Combine as metadata: .<table>Alter
49
Close database connection
Return
In most cases the initial position uncertainty of a particular dataset can be determined using:
• Metadata
• Technological parameters derived from metadata cross referenced to uncertainty parameters defined in:
• Academic papers
• Manuals
• etc.
The rule-of-thumb position error is dependent on the technology used to determine position by the platform at the time the elevation measurements were gathered.
This routine appears deceptively simple but should involve some research into the relevant position uncertainty. The results of the research should be held in a rule-of-thumb table to be cross referenced with the dataset capture date, platform, technology, etc.



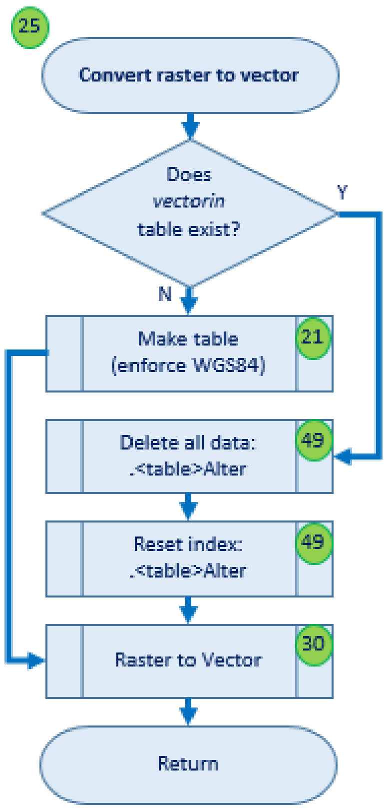




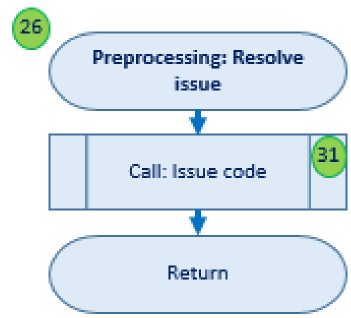
Import raster file into database.



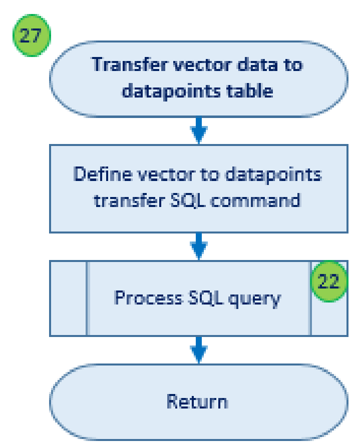
Create temp table then convert data.



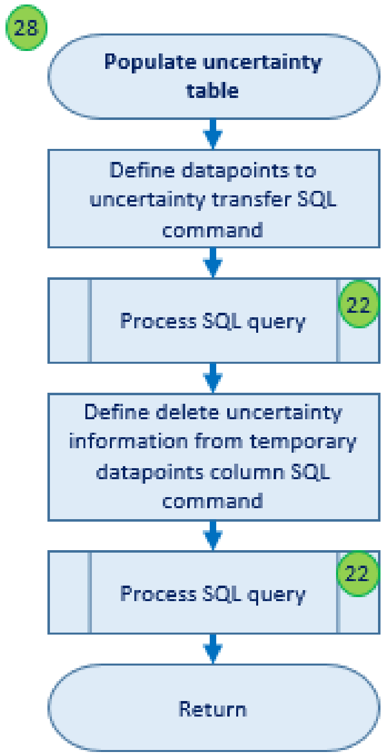
Issues are not necessarily predictable. Issue code should be in separate code files. Referenced by an ini file, the separate code will allow expansion of issue handling as required with minimal alteration to core code modules.

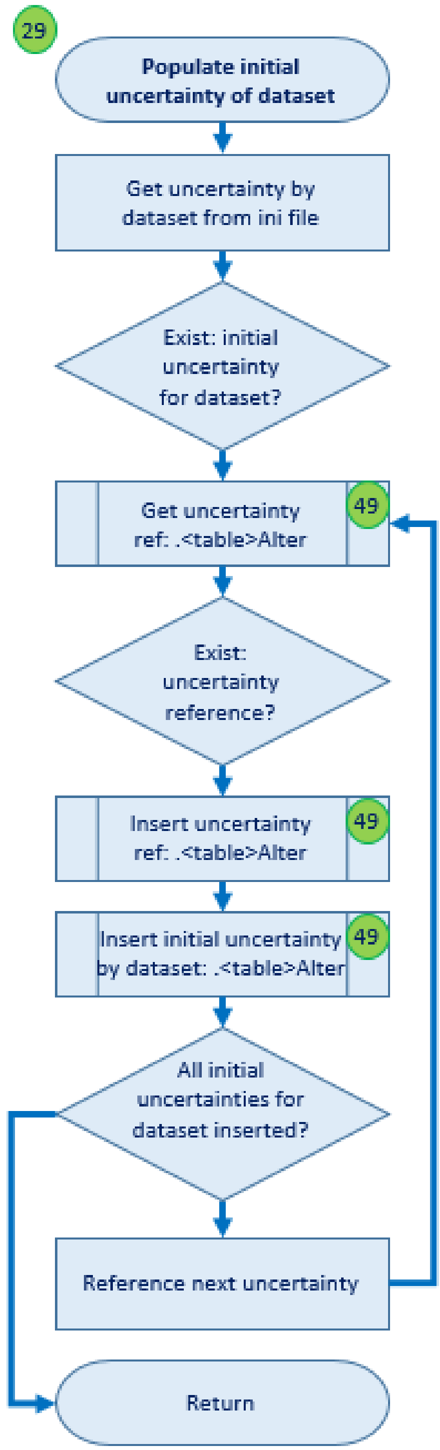


Transfer vector data from temporary processing table to permanent storage.

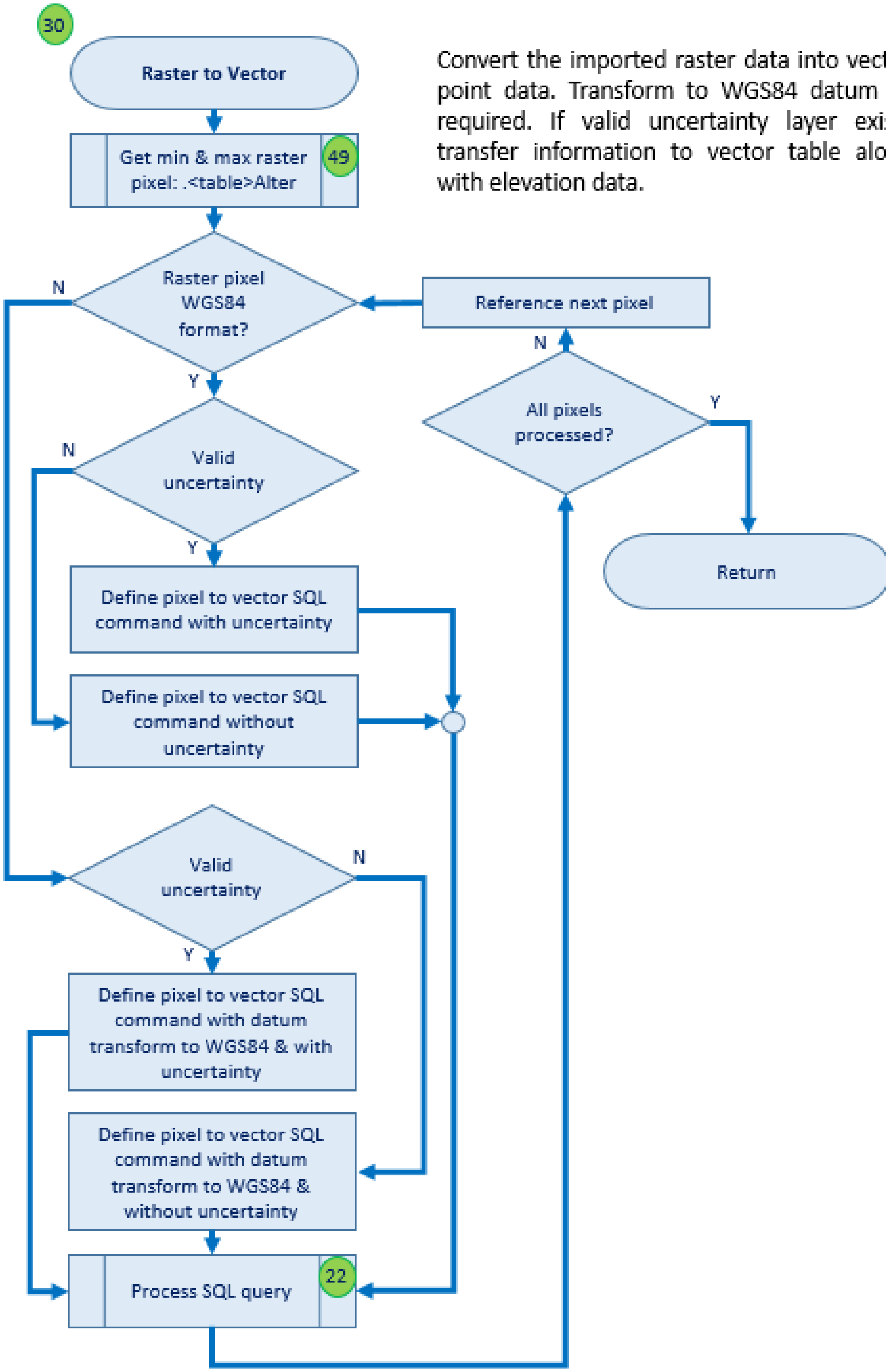


Transfer uncertainty data from temporary to permanent storage.

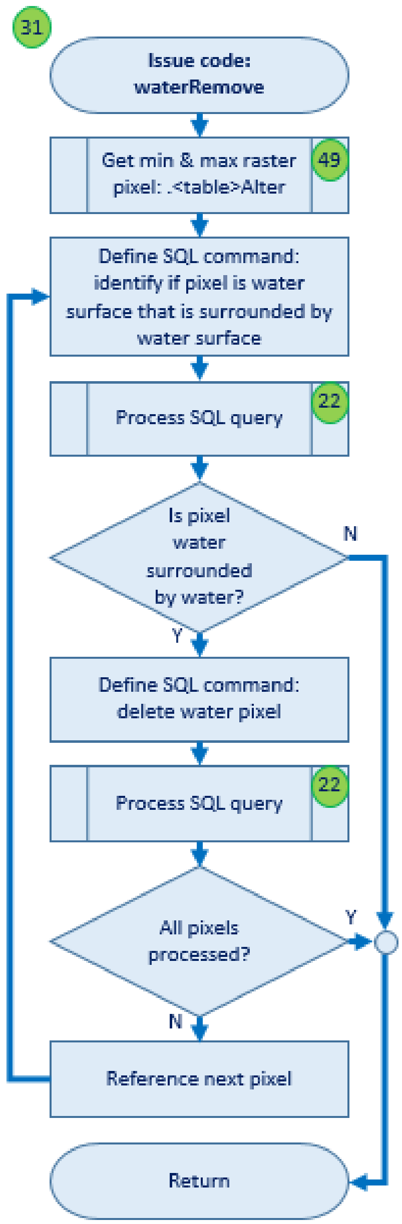




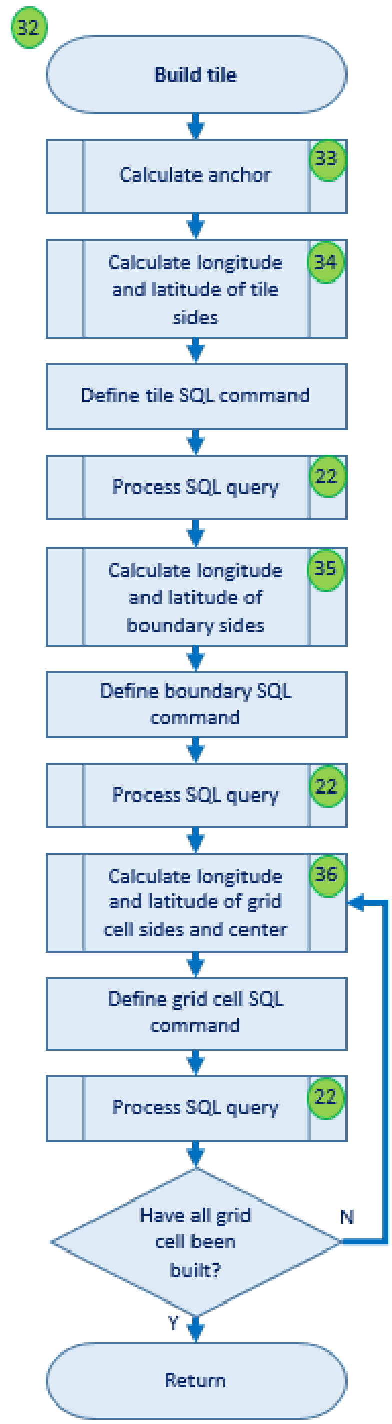
A dataset can have more than one initial uncertainty figure that applies to the whole dataset (e.g. average uncertainty). The various initial uncertainties are often derived from different sources that are referenced.

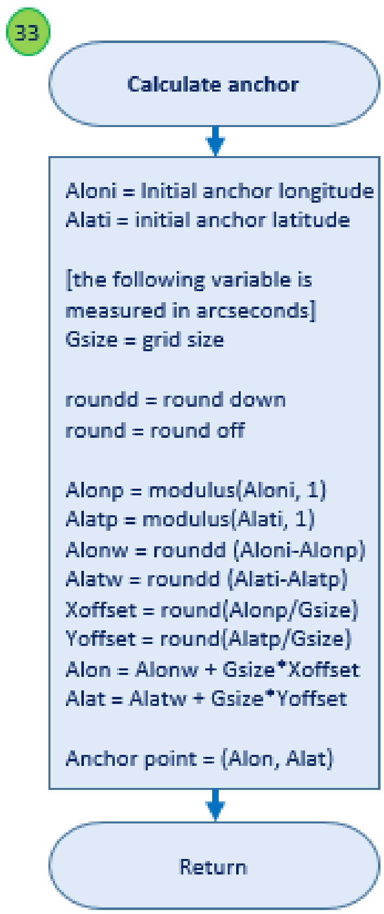


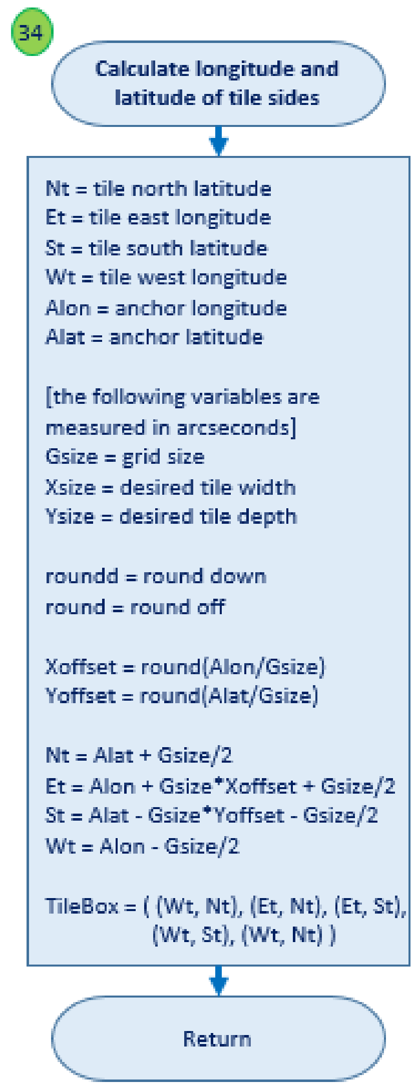
Convert the imported raster data into vector point data. Transform to WGS84 datum as required. If valid uncertainty layer exists transfer information to vector table along with elevation data.

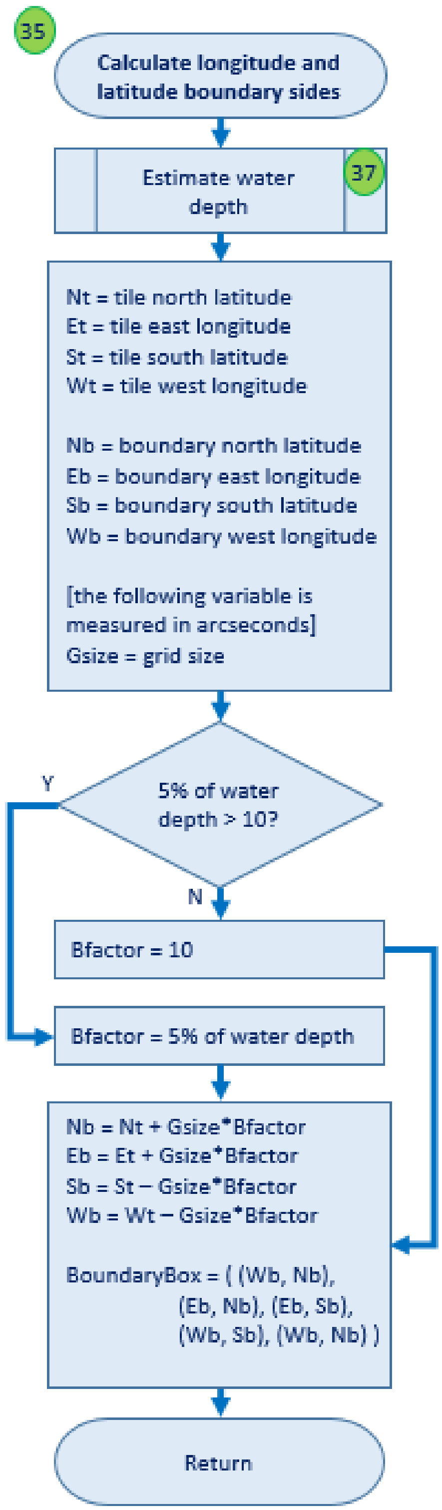


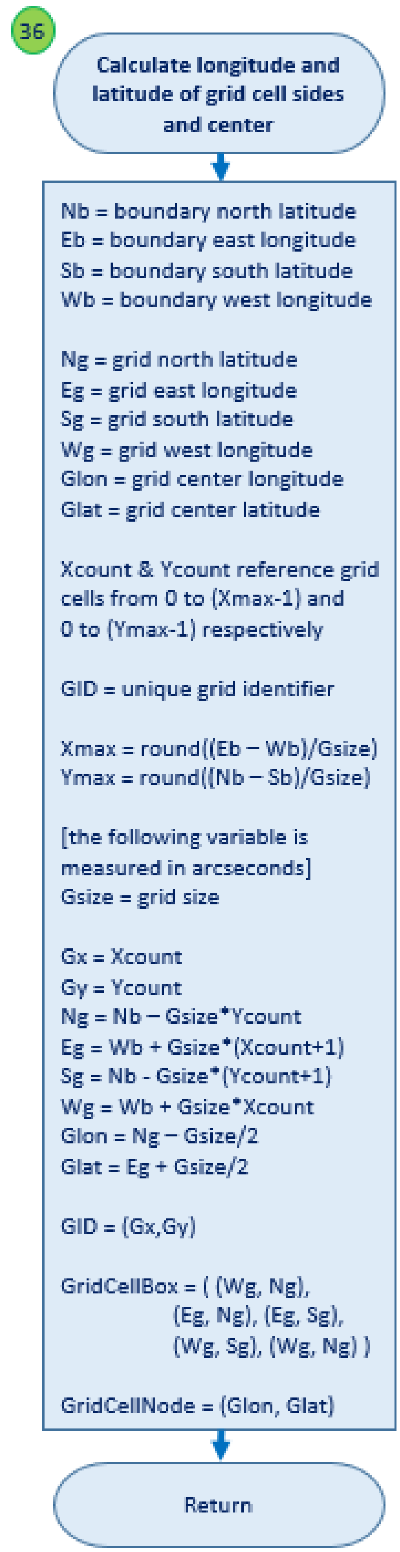
The water remove issue code is designed to delete all vector points corresponding to a raster pixel that is surrounded by water. This process level a zero elevation level border of points along all coastlines thereby preserving maximum terrain information.

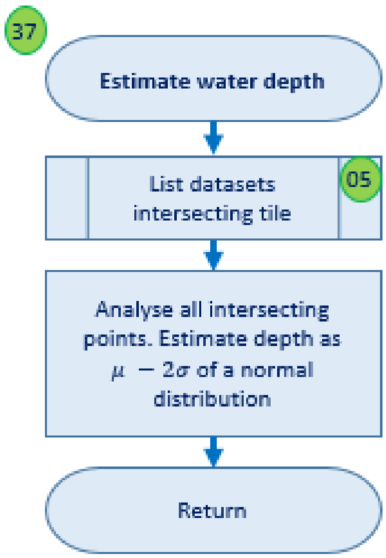



33
Calculate anchor
Aloni = Initial anchor longitude
Alati = initial anchor latitude
[the following variable is
measured in arcseconds]
Gsize = grid size
roundd = round down
round = round off
Alonp = modulus(Aloni, 1)
Alatp = modulus(Alati, 1)
Alonw = roundd (Aloni-Alonp)
Alatw = roundd (Alati-Alatp)
Xoffset = round(Alonp/Gsize)
Yoffset = round(Alatp/Gsize)
Alon = Alonw + Gsize*Xoffset
Alat = Alatw + Gsize*Yoffset
Anchor point = (Alon, Alat)
Return

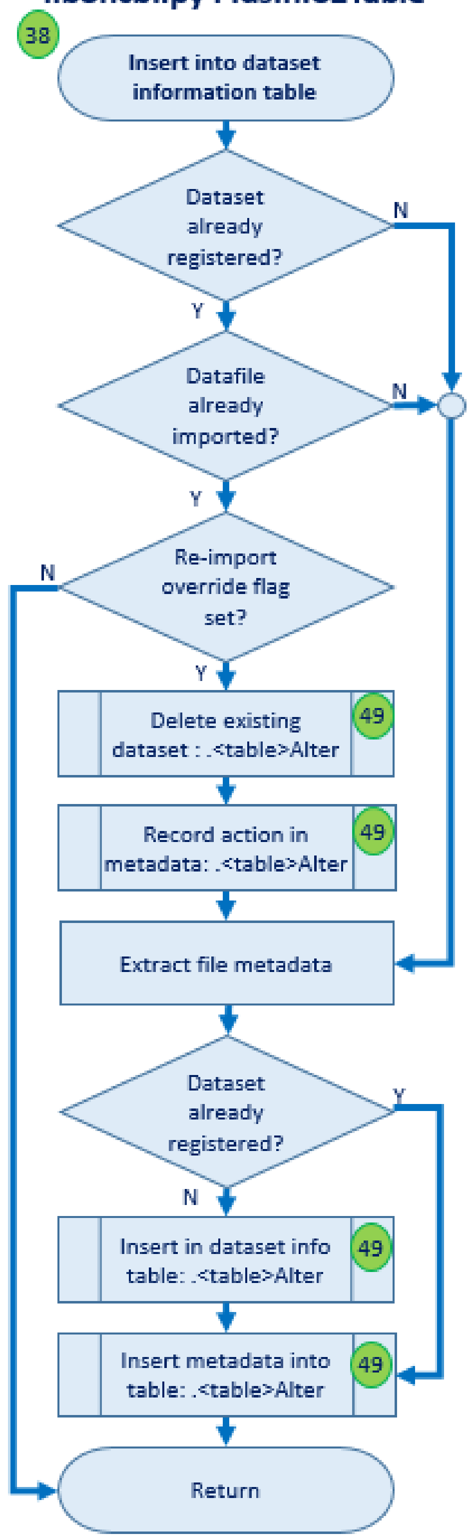

34
Calculate longitude and
latitude of tile sides
Nt = tile north latitude
Et = tile east longitude
St = tile south latitude
Wt = tile west longitude
Alon = anchor longitude
Alat = anchor latitude
[the following variables are
measured in arcseconds]
Gsize = grid size
Xsize = desired tile width
Ysize = desired tile depth
roundd = round down
round = round off
Xoffset = round(Alon/Gsize)
Yoffset = round(Alat/Gsize)
Nt = Alat + Gsize/2
Et = Alon + Gsize*Xoffset + Gsize/2
St = Alat - Gsize*Yoffset - Gsize/2
Wt = Alon - Gsize/2
TileBox = ((Wt, Nt), (Et, Nt), (Et, St),
(Wt, St), (Wt, Nt))
Return


35
Calculate longitude and latitude boundary sides
Estimate water depth
37
Nt = tile north latitude
Et = tile east longitude
St = tile south latitude
Wt = tile west longitude
Nb = boundary north latitude
Eb = boundary east longitude
Sb = boundary south latitude
Wb = boundary west longitude
[the following variable is measured in arcseconds]
Gsize = grid size
5% of water depth > 10?
Y
N
Bfactor = 10
Bfactor = 5% of water depth
Nb = Nt + Gsize*Bfactor
Eb = Et + Gsize*Bfactor
Sb = St – Gsize*Bfactor
Wb = Wt – Gsize*Bfactor
BoundaryBox = ((Wb, Nb), (Eb, Nb), (Eb, Sb), (Wb, Sb), (Wb, Nb))
Return


36
Calculate longitude and latitude of grid cell sides and center
Nb = boundary north latitude
Eb = boundary east longitude
Sb = boundary south latitude
Wb = boundary west longitude
Ng = grid north latitude
Eg = grid east longitude
Sg = grid south latitude
Wg = grid west longitude
Glon = grid center longitude
Glat = grid center latitude
Xcount & Ycount reference grid cells from 0 to (Xmax-1) and 0 to (Ymax-1) respectively
GID = unique grid identifier
Xmax = round((Eb – Wb)/Gsize)
Ymax = round((Nb – Sb)/Gsize)
[the following variable is measured in arcseconds]
Gsize = grid size
Gx = Xcount
Gy = Ycount
Ng = Nb – Gsize*Ycount
Eg = Wb + Gsize*(Xcount+1)
Sg = Nb - Gsize*(Ycount+1)
Wg = Wb + Gsize*Xcount
Glon = Ng – Gsize/2
Glat = Eg + Gsize/2
GID = (Gx,Gy)
GridCellBox = ((Wg, Ng), (Eg, Ng), (Eg, Sg), (Wg, Sg), (Wg, Ng))
GridCellNode = (Glon, Glat)
Return



liboncbf.py : .dsInfo2Table



A.4 Implemented Database Tables

These are the tables as implemented for the Data Homogenization Stage of iDEM in the first version of the prototype. These do not strictly adhere to the requirements of the flowcharts.

Please note that the tile database tables are included here but are not part of the Data Homogenization Stage of the prototype. These are currently under development. For clarity the following information should be noted [104][105]:

- ’varchar’: variable length string.
- ’integer’: integer data type with range -2147483648 to +2147483647.
- ’serial’: an autoincrementing integer data type.
- ’bigint’: large integer data type with a range -9223372036854775808 to 9223372036854775807.
- ’bigserial’: an autoincrementing large integer data.
- SRID `<code>`: code 4326 is a constraint that forces the location point to be defined in a longitude and latitude global coordinate system with WGS-84 datum; SRID name and authority are, respectively, the identifying handle (e.g. WGS-84) and the institution defining the standard.
- ’geographic point’: a PostGIS spacial data type that automatically prompts the spacial functionality to perform calculations such as distances between two geographic points as an arc on the surface of the SRID geoid instead of a straight line approximation.

datapoints table: stores the input data points

description	name	type	constraint	bytes
point identifier	pid	bigserial	PRIMARY KEY	8
dataset identifier	dsid	int		4
elevation value	elevation	double precision		8
lat, lon location	location	geography point	SRID 4326	48
processing buffer	adouble	double precison	DEFAULT 0	8
			total bytes	76

dsinfo table: information defining datasets

description	name	type	constraint	bytes
dataset identifier	dsid	serial		4
dataset name	dsname	varchar	PRIMARY KEY	variable
file name	dsfile	varchar	PRIMARY KEY	variable
SRID name	sridname	varchar		variable
SRID authority	sridauth	varchar		variable
SRID code	sridcode	varchar		variable
lat, lon box	boundry	geography polygon	SRID 4326	variable

dsmeta table: metadata about the datasets

description	name	type	constraint	bytes
metadata identifier	metaid	serial	PRIMARY KEY	8
dataset identifier	dsid	integer		4
event description	anevent	varchar		variable
event status	status	boolean		1
event detail	metastr	varchar		variable

somevariables table: miscellaneous information

description	name	type	constraint	bytes
variable name	varname	varchar	PRIMARY KEY	variable
variable value	varstring	varchar		variable

tiledatasets table: identifies datasets utilized in tile processing

description	name	type	constraint	bytes
tile identifier	tileid	integer	PRIMARY KEY	8
dataset identifier	dsid	integer	PRIMARY KEY	8

Table A.5: Database table definitions.

Table A.5: Database table definitions. (continued)

tileinfo table: defines tile geometry

description	name	type	constraint	bytes
tile identifier	tileid	serial	PRIMARY KEY	4
tile name	tilename	varchar		variable
lat, lon W,N corner reference	refpoint	geography point	SRID 4326	48
tile X size	tileX	double precison		8
tile Y size	tileY	double precison		8
grid cell size	gridsize	double precison		8
lat, lon box	tilebox	geography polygon	SRID 4326	variable

uncertref table: metadata - source of uncertainty information

description	name	type	constraint	bytes
uncertainty source identifier	refid	serial	PRIMARY KEY	8
uncertainty source string	uref	varchar		variable

uncertinitialbyds table: initial average uncertainty of a dataset

description	name	type	constraint	bytes
dataset identifier	dsid	bigint	PRIMARY KEY	8
uncertainty source identifier	refid	int	PRIMARY KEY	4
uncertainty x direction	ux	double precision		8
x uncertainty units	uxunit	varchar		variable
uncertainty y direction	uy	double precision		8
y uncertainty units	uyunit	varchar		variable
uncertainty z direction	uz	double precision		8
z uncertainty units	uzunit	varchar		variable

uncertinitialbypid table: initial uncertainty of a point

description	name	type	constraint	bytes
point identifier	pid	bigint	PRIMARY KEY	8
uncertainty source identified	refid	int	PRIMARY KEY	4
uncertainty x direction	ux	double precision		8
x uncertainty units	uxunit	varchar		variable
uncertainty y direction	uy	double precision		8
y uncertainty units	uyunit	varchar		variable
uncertainty z direction	uz	double precision		8
z uncertainty units	uzunit	varchar		variable

Table A.5: Database table definitions. (continues)

Table A.5: Database table definitions. (continued)

uncert table: combined uncertainty of a point

description	name	type	constraint	bytes
point identifier	pid	bigint	PRIMARY KEY	8
uncertainty estimation method	themethod	varchar	PRIMARY KEY	variable
grid cell size	gridsize	double precison	PRIMARY KEY	8
uncertainty value	uz	double precison		8
uncertainty units	uzunit	varchar		variable
assuming *themethod* 40 bytes and *uzunit* 11 bytes: total bytes				75

A.5 Initialization File

A.5.1 oncbf.ini

This initialization file currently serves a dual purpose in that it contains the database connection information as well as information about the input data. Input data that has not been defined in this configuration file can not be imported using the library. The need for this definition of the data encourages the user to become familiar with the data prior to importation by defining where input data can be located, which dataset it belongs to, the type or format of the data, average dataset uncertainty values, any known issues with the dataset that may need additional preprocessing. This should reduce the possibility of accidental importation of corrupt or inappropriate data. It will also add flexibility to the system being implemented by allowing importation to be partially customized.

As the prototype matures this initialization file will be divided into three separate files based on function:

- GDAL parameters [raster]
- database connection parameters [dbTarget]
- dataset configuration parameters [data]

```
[DEFAULT]

[raster]
tileX = 128
tileY = 128

[dbTarget]
dbName = onc
userID = user
pwd = password
host = localhost
port = 5432
schema = dev

[data]

chsbagFolder = data/CHSBAG/
chsbagExtension = .bag
chsbagType = raster
chsbagMultiFile = False
chsbagdsUref1 = Guesstimate by Peter Salomonsson
chsbagdsUx1 = 0
chsbagdsUxUnit1 = meter
chsbagdsUy1 = 0
chsbagdsUyUnit1 = meter
chsbagdsUz1 = 3.5
chsbagdsUzUnit1 = percent
chsbagissue1 = uncertcorupt

chsbaguFolder = data/CHSBAGu/
chsbaguExtension = .bag
chsbaguType = raster
chsbaguMultiFile = False

oncpaiFolder = data/oncPAI/
```

```
oncpaiExtension = .adf
oncpaiType = raster
oncpaiMultiFile = False

srtm1Folder = data/srtm1/
srtm1Extension = .hgt
srtm1Type = raster
srtm1MultiFile = True
srtm1dsUref1 = Rodriguez, Ernesto, et al. "An assessment of
               the SRTM topographic products." (2005).
srtm1dsUx1 = 12.6
srtm1dsUxUnit1 = meter
srtm1dsUy1 = 12.6
srtm1dsUyUnit1 = meter
srtm1dsUz1 = 9
srtm1dsUzUnit1 = meter
srtm1dsUref2 = Wagner, M. "SRTM DTED Format." Product Description
               SRTM/PD03/11/03, Version 1 (2003).
srtm1dsUx2 = 20
srtm1dsUxUnit2 = meter
srtm1dsUy2 = 20
srtm1dsUyUnit2 = meter
srtm1dsUz2 = 16
srtm1dsUzUnit2 = meter
srtm1issue1 = water

srtm3Folder = data/srtm3/
srtm3Extension = .hgt
srtm3Type = raster
srtm3MultiFile = True
srtm3dsUref1 = Rodriguez, Ernesto, et al. "An assessment of
               the SRTM topographic products." (2005).
srtm3dsUx1 = 12.6
srtm3dsUxUnit1 = meter
srtm3dsUy1 = 12.6
```

```
srtm3dsUyUnit1 = meter
srtm3dsUz1 = 9
srtm3dsUzUnit1 = meter
srtm3dsUref2 = Wagner, M. "SRTM DTED Format." Product Description
               SRTM/PD03/11/03, Version 1 (2003).
srtm3dsUx2 = 20
srtm3dsUxUnit2 = meter
srtm3dsUy2 = 20
srtm3dsUyUnit2 = meter
srtm3dsUz2 = 16
srtm3dsUzUnit2 = meter
srtm3issue1 = water

srtm30Folder = data/srtm30/srtm30/grd/
srtm30Extension = .nc
srtm30Type = raster
srtm30MultiFile = True
srtm30dsUref1 = Rodriguez, Ernesto, et al. "An assessment of
                the SRTM topographic products." (2005).
srtm30dsUx1 = 12.6
srtm30dsUxUnit1 = meter
srtm30dsUy1 = 12.6
srtm30dsUyUnit1 = meter
srtm30dsUz1 = 9
srtm30dsUzUnit1 = meter
srtm30dsUref2 = Wagner, M. "SRTM DTED Format." Product Description
                SRTM/PD03/11/03, Version 1 (2003).
srtm30dsUx2 = 20
srtm30dsUxUnit2 = meter
srtm30dsUy2 = 20
srtm30dsUyUnit2 = meter
srtm30dsUz2 = 16
srtm30dsUzUnit2 = meter
srtm30SRIDcode = 4326
```

Bibliography

[1] National Geospacial-Intelligence Agency. Office of geomatics: World geodetic system 1984 (wgs 84). http://earth-info.nga.mil/GandG/wgs84/. Accessed: 2014 11 26.

[2] Earth Science Australia. tsunamitidal waves. http://earthsci.org/education/teacher/basicgeol/tsumami/tsunami.html. Accessed: 2014 12 07.

[3] Samsuzana Abd Aziz. *Development of digital elevation models (DEMs) for agricultural applications*. ProQuest, 2008.

[4] Qimera Knowledge Base. World geodetic system 1984 (wgs84). https://confluence.qps.nl/pages/viewpage.action?pageId=29855173. Accessed: 2015 06 26.

[5] JJ Becker, DT Sandwell, WHF Smith, J Braud, B Binder, J Depner, D Fabre, J Factor, S Ingalls, SH Kim, et al. Global bathymetry and elevation data at 30 arc seconds resolution: Srtm30_plus. *Marine Geodesy*, 32(4):355–371, 2009.

[6] Javier Bello-Pineda and Jose Luis Hernández-Stefanoni. Comparing the performance of two spatial interpolation methods for creating a digital bathymetric model of the yucatan submerged platform. *Pan-American Journal of Aquatic Sciences*, 2(3):247–254, 2007.

[7] Geoff Bohling. Introduction to geostatistics and variogram analysis. *Kansas geological survey, 20p*, 2005.

[8] R. Briggs. Gis data structures. www.utdallas.edu/~briggs/poec5319/struct.ppt. Accessed: 2015 11 19.

[9] BR Calder and LA Mayer. Automatic processing of high-rate, high-density multibeam echosounder data. *Geochemistry, Geophysics, Geosystems*, 4(6), 2003.

[10] Brian Calder. On the uncertainty of archive hydrographic data sets. *Oceanic Engineering, IEEE Journal of*, 31(2):249–265, 2006.

[11] Brian Calder, Shannon Byrne, Bill Lamey, Richard T Brennan, James D Case, David Fabre, Barry Gallagher, R Wade Ladner, Friedhelm Moggert, and Mark Paton. The open navigation surface project. *The International hydrographic review*, 6(2), 2005.

[12] Brian Calder and David Wells. Cube users manual. *Unpublished report of the Center for Coastal and Ocean Mapping Group & Joint Hydrography Center, University of New Hamsphire, Durham, NH, USA*, 2007.

[13] Natural Resources Canada. Geospatial data extraction. `http://geogratis.gc.ca/site/eng/extraction`. Accessed: 2015 06 26.

[14] Natural Resources Canada. Map datum. `http://www.nrcan.gc.ca/earth-sciences/geography/topographic-information/maps/9791`. Accessed: 2014 11 27.

[15] B. Canova. Japan earthquake, tsunami: Moments of destruction. `http://abcnews.go.com/International/japan-earthquake-tsunami-moments-destruction/story?id=15862106`. Accessed: 2015 12 30.

[16] D.W. Caress and D.N. Chayes. Mb system, seafloor mapping software. `http://www.mbari.org/data/mbsystem/`. Accessed: 2015 12 21.

[17] J.H. Carter and B.C. Hadley et. al. *Topographic and Bathymetric Data Considerations: Datums, Datum Conversion Techniques, and Data Integration*. `http://coast.noaa.gov/digitalcoast/_/pdf/topographic-and-bathymetric-data-considerations.pdf`, 2007. Accessed: 2014 12 07.

[18] Center for Coastal and Ocean Mapping Joint Hydrographic Center. *CUBE and Navigation Surface: New Approaches for Hydrographic Data Processing and Management*, St. Andrews, New Brunswick, Canada, 2004.

[19] Center for Coastal and Ocean Mapping Joint Hydrographic Center. *Tackling Modern Multibeam Data with CUBE*, Hamburg, Hamburg, Germany, Nov 22 - Nov 25 2004.

[20] Pierre Cervenka, Christian De Moustier, and Peter F Lonsdale. Geometric corrections on sidescan sonar images based on bathymetry. application with seamarc ii and sea beam data. *Marine Geophysical Researches*, 16(5):365–383, 1994.

[21] William S Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American statistical association*, 74(368):829–836, 1979.

[22] William S Cleveland and Susan J Devlin. Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83(403):596–610, 1988.

[23] William S Cleveland, Susan J Devlin, and Eric Grosse. Regression by local fitting: methods, properties, and computational algorithms. *Journal of econometrics*, 37(1):87–114, 1988.

[24] Washington Department of Ecology CMAP. Glossary of coastal terminology. `http://www.ecy.wa.gov/programs/sea/swces/products/publications/glossary/words/A_C.htm`. Accessed: 2015 11 19.

[25] USGS Woods Hole Coastal and Marine Science Center. Bathymetry systems. `http://woodshole.er.usgs.gov/operations/sfmapping/bathyhist.htm`. Accessed: 2015 06 23.

[26] USGS Woods Hole Coastal and Marine Science Center. Sea-floor mapping technology, bathymetry systems. `http://woodshole.er.usgs.gov/operations/sfmapping/bathy.htm`. Accessed: 2015 06 23.

[27] USGS Woods Hole Coastal and Marine Science Center. Sidescan sonar systems. `http://woodshole.er.usgs.gov/operations/sfmapping/sonarhist.htm`. Accessed: 2015 06 23.

[28] USGS Woods Hole Coastal and Marine Science Center. Whsc bathymetry systems - single beam. http://woodshole.er.usgs.gov/operations/sfmapping/singlebeam.htm. Accessed: 2015 06 23.

[29] USGS Woods Hole Coastal and Marine Science Center. Whsc bathymetry systems - swath. http://woodshole.er.usgs.gov/operations/sfmapping/swath.htm. Accessed: 2015 06 23.

[30] DE Queiroz Cobra and T Daniel. Estimation and correction of geometric distortions in side-scan sonar images. Technical report, DTIC Document, 1990.

[31] Robert A Cohen. An introduction to proc loess for local regression. In *Proceedings of the 24th SAS users group international conference, Paper*, volume 273, 1999.

[32] O Conrad. Saga (system for automated geoscientific analyses), 2005.

[33] Helen Couclelis. People manipulate objects (but cultivate fields): beyond the raster-vector debate in gis. In *Theories and methods of spatio-temporal reasoning in geographic space*, pages 65–77. Springer, 1992.

[34] National Research Council. Technical issues in noaa's nautical chart program. https://books.google.ca/books?id=hFgrAAAAYAAJ.aspx. Accessed: 2015 11 19.

[35] Anna M Crawford. Methods of determining towfish location for improvement of side scan sonar image positioning. *DREA TM*, 19, 2002.

[36] B. de Lange Boom. *Vertical Datum Separation Models for the British Columbia Coast.* CHC 2012 The Arctic, Old Challenges New Approaches, 2012. Niagara Falls, Canada 15-17 May 2012.

[37] S Eberly, J Swall, D Holland, B Cox, and E Baldridge. Developing spatially interpolated surfaces and estimating uncertainty. *United States Environmental Protection Agency*, 2004.

[38] P.A. Elmore and S.J. Zambo. System and method for estimating uncertainty for geophysical gridding routines lacking inherent uncertainty estimation, September 12 2014. WO Patent App. PCT/US2014/021,486.

[39] Paul Elmore, David H Fabre, Raymond T Sawyer, R Wade Ladner, et al. Uncertainty estimation of historical bathymetric data from bayesian networks. In *OCEANS 2009, MTS/IEEE Biloxi-Marine Technology for Our Future: Global and Local Challenges*, pages 1–8. IEEE, 2009.

[40] Paul Elmore, Chad A Steed, et al. Algorithm design study for bathymetry fusion-review of current state-of-the-art and recommended design approach. Technical report, DTIC Document, 2008.

[41] esri. Arcgis for maritime: Bathymetry. `http://www.esri.com/software/arcgis/extensions/maritime/bathymetry/features`. Accessed: 2015 11 19.

[42] Fisheries and Ocean Canada. Herring spawn and catch records - bc search map. `http://www.pac.dfo-mpo.gc.ca/science/species-especes/pelagic-pelagique/herring-hareng/herspawn/pages/default0-eng.html`. Accessed: 2015 04 12.

[43] Fisheries and Oceans Canada. Hydrography online - data acquisition - depths and heights - measuring depths with echosounders and sonars. `http://www.dfo-mpo.gc.ca/science/hydrography-hydrographie/measuring-depths-eng.html`. Accessed: 2014 11 14.

[44] Fisheries and Oceans Canada. Using lasers to measure depths and heights. `http://www.dfo-mpo.gc.ca/science/hydrography-hydrographie/lasers-eng.html`. Accessed: 2014 12 07.

[45] Kenneth E Foote and Donald J Huebner. Error, accuracy, and precision. *The Geographer's Craft*, 1995.

[46] Center for Coastal and Ocean Mapping. Cube. `http://ccom.unh.edu/theme/data-processing/cube`. Accessed: 2014 11 22.

[47] MGG Foreman, WR Crawford, JY Cherniawsky, RF Henry, and MR Tarbotton. A high-resolution assimilating tidal model for the northeast pacific ocean. *Journal of Geophysical Research: Oceans (1978–2012)*, 105(C12):28629–28651, 2000.

[48] J Gallant and J Austin. Stitching fine resolution dems. In *18th World IMACS Congress and MODSIM09 International Congress on Modelling and Simulation*, pages 2486–2492, 2009.

[49] GDAL. Geospatial data abstraction library. http://www.gdal.org/. Accessed: 2015 05 02.

[50] GEBCO. General bathymetric chart of the oceans. http://www.gebco.net/data_and_products/gridded_bathymetry_data/. Accessed: 2015 06 24.

[51] GMV. Glonass performances. http://www.navipedia.net/index.php/GLONASS_Performances. Accessed: 2015 11 19.

[52] A Goodwillie. User guide to the gebco one minute grid. *Centenary edition of the GEBCO digital atlas, GEBCO*, 2003.

[53] Australian Government and SLSA Training Resource. Tsunami education and awareness. http://m.beachsafe.org.au/tsunami. Accessed: 2015 04 12.

[54] Greg Grohman, George Kroenung, and John Strebeck. Filling srtm voids: the delta surface fill method. *Photogrammetric Engineering and Remote Sensing*, 72(3):213–216, 2006.

[55] Harvard. A summary of error propagation. http://ipl.physics.harvard.edu/wp-uploads/2013/03/PS3_Error_Propagation_sp13.pdf, 2007. Accessed: 2015 06 18.

[56] K Hayat, W Puech, and G Gesquiere. Seamless heterogeneous tessellation via smoothing and mosaicking in the dwt domain. In *IS&T/SPIE Electronic Imaging*, pages 75430W–75430W. International Society for Optics and Photonics, 2010.

[57] Benjamin Hell and Martin Jakobsson. Gridding heterogeneous bathymetric data sets with stacked continuous curvature splines in tension. *Marine Geophysical Research*, 32(4):493–501, 2011.

[58] Tianhang Hou, Lloyd C Huff, and Larry Mayer. Automatic detection of outliers in multibeam echo sounding data. *University of New Hampshire*, pages 1–12, 2001.

[59] Tianhang Hou, Lloyd C Huff, and Larry Mayer. Automatic detection of outliers in multibeam echo sounding data. *University of New Hampshire*, pages 1–12, 2001.

[60] QPS hydrographic and marine software solutions. Cube algorithm. `https://confluence.qps.nl/display/KBE/CUBE+Algorithm`. Accessed: 2014 12 04.

[61] IOC IHO. The iho-ioc gebco cookbook. `http://www.star.nesdis.noaa.gov/sod/lsa/GEBCO_Cookbook/documents/CookBook.9.3.14.pdf`. Accessed: 2014 11 25.

[62] Google images. 20041230 tsunami form graphic.gif. `https://acccearthscience.wikispaces.com/Formation+of+Tsunamis+-+SO`. Accessed: 2014 12 07.

[63] Google images. 63409 wl 91949v00 sonarscope fig2. `http://www.mathworks.com/company/newsletters/articles/developing-and-deploying-sonar-and-echosounder-data-analysis-software.html`. Accessed: 2014 12 07.

[64] Google images. Bc utm zones. `http://www.empr.gov.bc.ca/MINING/GEOSCIENCE/MAPPLACE/MAINMAPS/Pages/BCUTMZones.aspx`. Accessed: 2014 11 27.

[65] Google images. History of hydrographic surveying. `http://www.nauticalcharts.noaa.gov/mcd/learnnc_surveytechniques.html`. Accessed: 2014 12 07.

[66] Google images. How to find pixel coordinates of a city on a world map using logitude and latitude? `http://www.satsig.net/lat_long.htm`. Accessed: 2014 11 27.

[67] Google images. How tsunamis form. `https://acccearthscience.wikispaces.com/Formation+of+Tsunamis+-+SO`. Accessed: 2014 12 07.

[68] Google images. Mbari - auv: Sonar system. http://tighar.org/Projects/Earhart/Niku7/niku7expedition.html. Accessed: 2014 12 07.

[69] Google images. Multibeam systems - sound velocity. http://www.amloceanographic.com/CTD-Sound-Velocity-Environmental-Instrumentation-Home/Multibeam-Overvew. Accessed: 2014 12 07.

[70] Google Images. Our products: Scalgo model scal. http://scalgo.com/solutions/product-model.php. Accessed: 2015 07 03.

[71] Google images. Preprocessing. http://foto.hut.fi/research/Strategy_98/Academic_suite/Reconstruction_procedure/Examples/Quide_map_Petra/Preprocessing/preprocessing.html. Accessed: 2014 12 07.

[72] Google images. Reality measurements inc. close range digital photogrammetry. http://www.realitymeasurements.com/crdp-low-altitude.htm. Accessed: 2014 12 07.

[73] Google images. Service innovation - ..ro corporate web. http://www.gsiseabed.ie/news02.htm. Accessed: 2014 12 07.

[74] Google images. Side scan sonar - underwater searching. http://gralston1.home.mindspring.com/Sidescan.html. Accessed: 2015 06 23.

[75] Google images. Space in images - 2012 - 05 - mapping. http://www.gebco.net/general_interest/faq/. Accessed: 2015 06 23.

[76] Google images. zego.co.nz. http://geophysicistatlarge.blogspot.ca/2014/01/seismic-data-orkney.html. Accessed: 2014 12 07.

[77] Tania Insua, Annette Grilli, Stephan Grilli, Michael Shelby, Kelin Wand, Dawei Gao, Josef Cherniawsky, Harris Jeffrey, Martin Heesemann, Scott McLean, and Kate Moran. Preliminary tsunami hazard assessment in british columbia, canada. AGU Fall Meeting 2015. NH23C-1890.

[78] Proc J. Hyperbolic navigation systems. http://www.jproc.ca/hyperbolic/. Accessed: 2015 11 19.

[79] Martin Jakobsson, Brian Calder, and Larry Mayer. On the effect of random errors in gridded bathymetric compilations. *Journal of Geophysical Research: Solid Earth (1978–2012)*, 107(B12):ETG–14, 2002.

[80] Martin Jakobsson, Brian Calder, Larry Mayer, and Andy Armstrong. The uncertainty of a bathymetric contour: Implications for the cut-off line. In *Acuracies and Uncertainties in Maritime Boundaries and Outer Limits, ABLOS Conference*. Citeseer, 2001.

[81] Martin Jakobsson, Ron Macnab, Larry Mayer, Robert Anderson, Margo Edwards, Jörn Hatzky, Hans Werner Schenke, and Paul Johnson. An improved bathymetric portrayal of the arctic ocean: Implications for ocean modeling and geological, geophysical and oceanographic analyses. *Geophysical Research Letters*, 35(7), 2008.

[82] S. James. What is the difference between vector and raster data models. `http://gis.stackexchange.com/questions/57142/what-is-the-difference-between-vector-and-raster-data-models`. Accessed: 2015 11 19.

[83] NC Kelland et al. Accurate acoustic position monitoring of deepwater geophysical towfish. In *Offshore Technology Conference*. Offshore Technology Conference, 1988.

[84] AM Coleman KB Larson, DL Ward, and JW Lettrick. Development of a high-resolution bathymetry dataset for the columbia river through the hanford reach. 2010.

[85] CB Lirakis and KP Bongiovanni. Automated multibeam data cleaning and target detection. In *OCEANS 2000 MTS/IEEE Conference and Exhibition*, volume 1, pages 719–723. IEEE, 2000.

[86] Lubos Mitas and Helena Mitasova. Spatial interpolation. *Geographical information systems: principles, techniques, management and applications*, 1:481–492, 1999.

[87] Karen A Mulcahy and Keith C Clarke. Symbolization of map projection distortion: a review. *Cartography and geographic information science*, 28(3):167–182, 2001.

[88] Mary Natrella. Nist/sematech e-handbook of statistical methods. 2010.

[89] GPS Navstar. *Global Positioning System Standard Positioning Service Signal Specification*. `http://www.gps.gov/technical/ps/1993-SPS-signal-specification.pdf`, 1st edition edition, 11 1993. Accessed: 2015 07 21.

[90] GPS Navstar. *Global Positioning System Precise Positioning Service Signal Specification*. `http://www.gps.gov/technical/ps/2007-PPS-performance-standard.pdf`, 2 2007. Accessed: 2015 07 21.

[91] S. Nelson. Natural disasters: Tsunami. `http://www.tulane.edu/~sanelson/Natural_Disasters/tsunami.htm`, 07 2012. Accessed: 2014 12 07.

[92] NOAA. Elementary edu noaa's national ocean and atmospheric administration. `http://oceanservice.noaa.gov/education/seafloor-mapping/how_history.html`. Accessed: 2015 06 23.

[93] NOAA. Multibeam bathymetry. `https://www.ngdc.noaa.gov/mgg/bathymetry/multibeam.html`. Accessed: 2015 12 01.

[94] NOAA. National geodetic survey faq. `http://www.ngs.noaa.gov/faq.shtml#WhatNAD`. Accessed: 2015 11 19.

[95] National Oceanic and Atmospheric Administration. Bathymetric data viewer. `http://maps.ngdc.noaa.gov/viewers/bathymetry/`. Accessed: 2014 11 24.

[96] NOAA Office of Coast Survey. Lidar. `http://www.nauticalcharts.noaa.gov/hsd/lidar.html`. Accessed: 2014 12 07.

[97] NOAA Office of Coast Survey. Side scan sonar. `http://www.nauticalcharts.noaa.gov/hsd/SSS.html`. Accessed: 2015 06 23.

[98] Minnesota Department of Natural Resources. Dem introduction to rasters and vectors. http://www.dnr.state.mn.us/watersheds/dem.html. Accessed: 2015 11 19.

[99] Scripps Institution of Oceanography. Global topography. http://topex.ucsd.edu/WWW_html/srtm30_plus.html. Accessed: 2014 11 21.

[100] University of the Western Cape. Inverse distance weighting (idw). http://planet.botany.uwc.ac.za/nisl/GIS/spatial/chap_1_14.htm. Accessed: 2015 07 02.

[101] Fugro Pelagos. Technical information on lidar bathymetry. http://www.fugro-pelagos.com/lidar/tech/lidar_bathy.html. Accessed: 2015 11 19.

[102] PennState. Geog482 ellipsoids. https://www.e-education.psu.edu/geog482spring2/c2_p15.html. Accessed: 2015 11 19.

[103] Raju P.L.N. Fundamentals of geographical information systems. http://wamis.org/agm/pubs/agm8/Paper-6.pdf. Accessed: 2015 11 19.

[104] PostGIS. Postgis. http://postgis.net/. Accessed: 2015 04 12.

[105] PostgreSQL. Postgresql. http://www.postgresql.org/. Accessed: 2015 04 12.

[106] WAAS PS2008. Global positioning system wide area augmentation system (waas) performance standard. *Washington DC: FAA*, 2008.

[107] python. python. http://www.python.org. Accessed: 2015 04 12.

[108] Spacial Reference. Epsg:4326 wgs 84. http://spatialreference.org/ref/epsg/wgs-84/. Accessed: 2014 11 27.

[109] CHS Rob Hare, Barry Eakins, and CIRES Chris Amante. Modeling bathymetric uncertainty. *US HYDRO 2011*, 2001.

[110] Catherine Robin and et. al. *The Continuous Vertical Datum for Canadian Waters Project: Status report and update.* EGU General Assembly Conference, 2014, 2014.

[111] Ernesto Rodriguez, CS Morris, JE Belz, EC Chapin, JM Martin, W Daffer, and S Hensley. An assessment of the srtm topographic products. 2005.

[112] William BF Ryan, Suzanne M Carbotte, Justin O Coplan, Suzanne O'Hara, Andrew Melkonian, Robert Arko, Rose Anne Weissel, Vicki Ferrini, Andrew Goodwillie, Frank Nitsche, et al. Global multi-resolution topography synthesis. *Geochemistry, Geophysics, Geosystems*, 10(3), 2009.

[113] Movable Type Scipts. Calculate distance, bearing and more between latitude/longitude points. `http://www.movable-type.co.uk/scripts/latlong.html`. Accessed: 2015 05 02.

[114] FM Seifert. A data fusion algorithm for dem mosaicking: Building a global dem with srtm-x and ers data. 2006.

[115] Canadian Hydrographic Service. Cube bathymetric data processing and analysis. `http://www.charts.gc.ca/data-gestion/hydrographic/bathymetric-traitement-eng.pdf`. Accessed: 2014 12 7.

[116] Relief Shading. Dem file formats. `http://www.reliefshading.com/analytical/dem-file-formats/`. Accessed: 2015 11 19.

[117] Walter HF Smith and David T Sandwell. Conventional bathymetry, bathymetry from space, and geodetic altimetry. *OCEANOGRAPHY-WASHINGTON DC-OCEANOGRAPHY SOCIETY-*, 17(1):8–23, 2004.

[118] WHF Smith and P Wessel. Gridding with continuous curvature splines in tension. *Geophysics*, 55(3):293–305, 1990.

[119] Tetushi Tachikawa, Manabu Kaku, Akira Iwasaki, Dean B Gesch, Michael J Oimoen, Zheng Zhang, Jeffrey J Danielson, Tabatha Krieger, Bill Curtis, Jeff Haase, et al. Aster global digital elevation model version 2-summary of validation results. Technical report, NASA, 2011.

[120] Marc Véronneau, Robert Duval, and Jianliang Huang. A gravimetric geoid model as a vertical datum in canada. *Geomatica*, 60(2):165–172, 2006.

[121] Slobodan Vucetic, Tim Fiez, and Zoran Obradovic. Examination of the influence of data aggregation and sampling density on spatial estimation. *Water Resources Research*, 36(12):3721–3730, 2000.

[122] M Wagner. Srtm dted format. *Product Description SRTM/PD03/11/03, Version*, 1, 2003.

[123] Robert Ward, Lee Alexander, and Barrie Greenslade. Iho s-100: The new iho hydrographic geospatial standard for marine data and information. *The International Hydrographic Review*, (1), 2009.

[124] Frank Warmerdam. The geospatial data abstraction library. In *Open Source Approaches in Spatial Data Handling*, pages 87–104. Springer, 2008.

[125] Paul Wessel, Walter HF Smith, Remko Scharroo, Joaquim Luis, and Florian Wobbe. Generic mapping tools: improved version released. *Eos, Transactions American Geophysical Union*, 94(45):409–410, 2013.

[126] Chin-Shung Yang, Szu-Pyng Kao, Fen-Bin Lee, and Pen-Shan Hung. Twelve different interpolation methods: A case study of surfer 8.0. In *Proceedings of the XXth ISPRS Congress*, volume 35, pages 778–785, 2004.

[127] Fang Yao. *Uncertainty analysis on photogrammetry-derived national shoreline*. PhD thesis, University of New Hampshire, 2014.

[128] Chunmei Zhao, Jikun Ou, and Yunbin Yuan. Positioning accuracy and reliability of galileo, integrated gps-galileo system based on single positioning model. *Chinese Science Bulletin*, 50(12):1252–1260, 2005.