International Journal of Geo-Information

MDPI

Article

Exploring the Use of Data in a Digital Twin for the Marine and Coastal Environment

Shelley Haupt [1,*], Bolelang Sibolla [1,2], Raymond Molapo [1,3], Lizwe Mdakane [1] and Nicolene Fourie [1]

[1] Spatial Information Systems, Next Generation Enterprises and Institutions, Council for Scientific and Industrial Research, Pretoria 0001, South Africa; bsibolla@csir.co.za (B.S.); rmolapo@csir.co.za (R.M.); lmdakane@csir.co.za (L.M.); nfourie@csir.co.za (N.F.)
[2] Department of Geography, Geoinformatics and Meteorology, University of Pretoria, Pretoria 0028, South Africa
[3] Multilingual Speech Technologies, North-West University, Potchefstroom 2520, South Africa
* Correspondence: shaupt@csir.co.za

Abstract: The ocean plays a vital role in our society and represents a constantly changing landscape that is not well understood and therefore needs continuous monitoring and research. Sustainable monitoring is essential to assess both the current and future state of our oceans. However, conventional monitoring faces significant challenges, including issues of accessibility, and spatial and temporal constraints. The development of digital twins of the ocean (DTO) offers an emerging technology that could revolutionise our understanding of marine and coastal environments. Current DTO have shown effectiveness in monitoring marine and coastal environments in the European context. However, there is a need for a DTO for the Southern African and Western Indian Ocean regions that addresses specific concerns that are relevant to these regions. Successful development of a DTO depends on the availability of high-quality data. Therefore, various data inputs are necessary to build an accurate digital twin. This paper explores the data that can be utilised in a DTO, detailing how different ocean variables are collected and integrated into the digital twin. As a first step towards the development of a DTO in these regions, the paper proposes a data management plan and its implementation in the development of DTO. The data management plan is based on the phases of data in a geospatial data life cycle. Challenges regarding the management of data in this DTO and possible solutions are presented in the conclusion.

Keywords: digital twin; ocean monitoring; marine data sources; data management plan; geospatial data life cycle

Academic Editors: Wolfgang Kainz and Dev Raj Paudyal

Received: 19 November 2024
Revised: 14 March 2025
Accepted: 20 March 2025
Published: 25 March 2025

Citation: Haupt, S.; Sibolla, B.; Molapo, R.; Mdakane, L.; Fourie, N. Exploring the Use of Data in a Digital Twin for the Marine and Coastal Environment. *ISPRS Int. J. Geo-Inf.* **2025**, *14*, 140. https://doi.org/10.3390/ijgi14040140

1. Introduction

The oceans cover over 70% of the Earth's surface and form a fundamental part of human and marine life [1]. They are recognised as one of humanity's most important natural resources [2,3]. Marine economies have experienced remarkable growth, providing several resources such as food, minerals, shipping, and water [4]. According to Teh et al. [5] millions of livelihoods are dependent on the global fisheries industry. Furthermore, the ocean also plays a vital role in regulating the Earth's climate by absorbing 30% of the carbon dioxide that is produced by human activities [3]. Despite the importance of the ocean, the sustainability of the marine ecosystem in Southern Africa and the Western Indian Ocean region is under threat, as a result of pollution from various invasive organisms and materials [3]. According to global plastics production statistics, approximately 0.5%

of plastic waste ends up in the ocean [6]. Therefore, it is important to conduct frequent monitoring to evaluate the current and future state of the oceans. However, there are a few challenges presented when monitoring due to the large scale of our oceans. These include complex currents and tides that can cause shifts in monitoring ocean parameters such as salinity [7]. Conventional techniques include in situ monitoring instrumentation such as physical and chemical based sensors, buoys, and underwater platforms [7,8]. These approaches rely on sampling at discrete points in the ocean, which is often time consuming and fails to provide information over changes in time [7]. Therefore, there is a need for an open and digital representation of the oceans which can provide necessary available data sources for sustainable ocean monitoring in a single platform, in near real time [8].

Recent advancements in digital technologies have led to a wide range of opportunities to improve ocean sustainability. One of these opportunities is the concept of a digital twin of the ocean (DTO). This specifically refers to the utilisation of physical models, sensors, operational data and additional relevant information to simulate multidisciplinary and multiscale processes [8–10]. The aim of the DTO is to create a virtual representation that can mirror the entire life cycle of its physical entity [9].

Real-world monitoring data can be used with computing technologies to simulate the behaviour of a physical entity [11]. The concept of a digital twin was first introduced as an 'information mirroring model' [12]. With the aid of emerging technologies, more industries have adopted the concept of digital twins. The European Commission is leading the development of the digital twin of the ocean through its Destination Earth (DE) and Digital Twins of the Ocean (DITTO) projects in alignment with the United Nations Ocean Decade programme [13]. Coupled with advanced technologies such as the Internet of Things (IoT), deep learning, cloud computing and artificial intelligence (AI), a DTO provides several benefits to the marine industry [4]. To date, several digital twin technologies in the marine and coastal environments have been developed [11,14,15]. The ILIAD project was developed to operate and showcase a set of DTOs which will support the design, development, and operation of innovative services related to the oceans and seas [14]. Through this project, a high-resolution digital twin pilot for the Cretan Sea was developed which focuses on oil spill pollution monitoring. The aim of the digital twin is to aid the immediate response in case of any accidental oil releases [16]. Another example is a digital twin of the North Sea that focusses on the improvement of satellite-based monitoring of essential coastal variables for the coastal region [8]. Through these technologies, various ocean parameters such as coastal zone spatial boundaries, water quality and aquaculture, oil spills, marine pollution, and ocean energy potential are monitored. The DTO provides an environment where these outputs can be accessed and analysed by multiple stakeholders, to promote and facilitate communication and collaboration [17].

Several technology frameworks have been developed for digital twins, in general, including Tao and Zhang [18], who proposed a four-dimensional framework of the digital twin which includes the following parameters: a physical entity, the data of the DT, virtual models, and services (Figure 1).

These general frameworks are further aligned and expanded to the DTO by [9]. The physical entity refers to the real object of the digital twin [18]. The virtual model refers to the representation of the physical entity in cyberspace, whilst algorithms and models within the digital twin facilitate further analysis and provide more understanding of the physical entity. The data dimension of the digital twin contains the data sources as well as the data management plan that is used to capture and describe the procedures on how to handle data throughout their life cycles. Singh et al. [19] further describe the information flow in a digital twin, by segmenting the digital twin into three layers. The first layer is the physical layer, which describes the configuration of sensors that facilitates data collection

in the physical space. The second layer is the data layer, which defines and describes how acquired data are received and packaged to become information that can be further analysed or stored for future analyses. The third and final layer is the model layer which describes the models that are used to transform data into information to provide further insights about the physical entity.

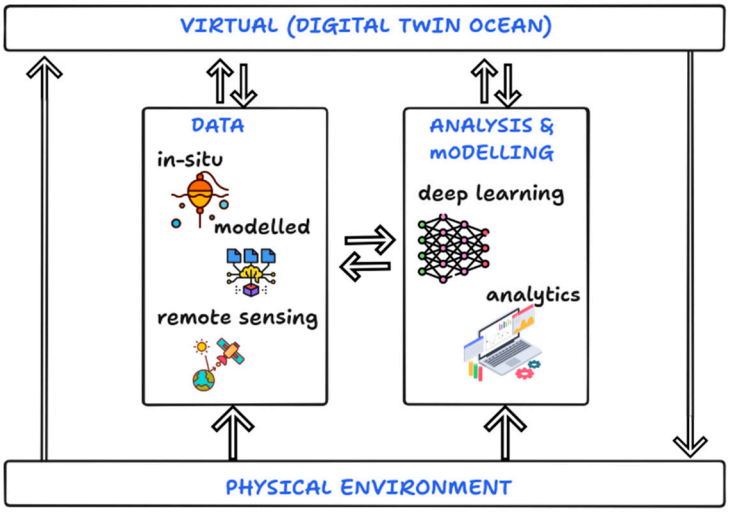

Figure 1. A DT framework and the interaction between the four-dimensional components.

Based on the various frameworks, it is evident that data plays a significant role in the successful development of a digital twin. Consequently, the development and successful implementation of a DTO relies on the availability of high-quality data [10]. The data are important for constructing virtual models, creating virtual connections between the physical entities and executing intelligent operations [9]. Combining the data sources above can enhance the accuracy, efficiency, and adaptability of DT-based services [10].

There has been limited focus on development and application of digital twins in African oceans. A DTO in this region can help provide valuable insights into the current state as well as future projections of both the ocean ecosystems and the human activities that impact them. Therefore, the development of a digital twin in the Southern African and Western Indian Ocean regions is important for maritime research, climate change preparedness and the management of the blue economy. The Southern Africa and Western Indian Ocean regions are shown Figure 2. These regions include coastal Southern and Western Indian Ocean countries in Africa. The countries included are Angola, Namibia, South Africa, Mozambique, Tanzania, Kenya, Madagascar, and the small islands, including Comoros, Seychelles and Mauritius which are shown is outlined in Figure 2.

As a first step towards the development and implementation of a digital twin of the ocean for Southern African and Western Indian Ocean regions, this paper aims to evaluate the various data requirements for a DTO by listing all the different data sources and outlining their advantages and disadvantages. Furthermore, the paper provides an overview of the data management plan and suggests a relevant implementation thereof towards its inclusion in a digital twin of the Southern African and Western Indian Ocean regions. The paper is organised as follows: Section 2 provides an overview of the role of data in a digital twin of the ocean. Firstly, the data considerations are outlined, and then common data types and collection methods are presented. In Section 3, current data management practises for digital twins are discussed, and the data life cycle is described. Following a description of the data life cycle, Section 4 discusses various data management principles that can be used as a guide for the effective management of the data throughout the data life cycle. Section 5 presents the challenges regarding data management considering the data

sources outlined in this paper and proposes further work. Finally, in Section 6, concluding remarks are presented.

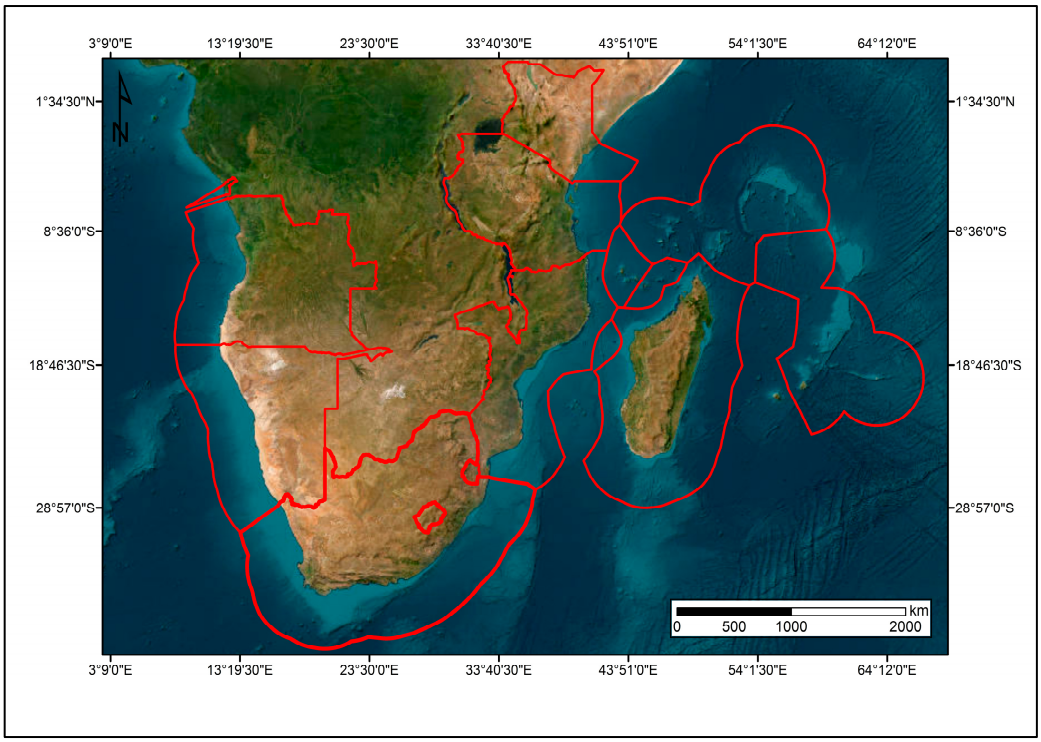

Figure 2. South African and Western Indian Ocean regions.

2. Data in a Digital Twin of the Ocean (DTO)

A basic DTO system consists of a physical entity and a virtual model connected by two-way data flow, enabling advanced analysis and continuous improvement [20]. The data that are used in a digital twin are driven by the specific needs and objectives of end users. Based on the literature, there are several data considerations that need to be taken into account based on user and application requirements [3,10,20,21]. These considerations are depicted in Figure 3 and discussed as follows. The first consideration is the data collection method, which refers to how the data are acquired or collected. The collected data can further be categorised into primary and secondary data. Primary data refers to data that are derived directly from the source. Secondary data refers to data that are derived from primary data, for example, spectral indices derived from satellite data. The second consideration is the accessibility and availability of data. The accessibility of data is the ease with which a user can find and retrieve data, whilst data availability measures how frequently data can be accessed by the user. The third consideration is the quantity and frequency. The quantity of the data refers to how much data can be retrieved, and the frequency refers to how often the data are available. The fourth and final consideration is the concept of data reliability. This refers to the completeness, consistency, and accuracy of data. Ideally, the more reliable the data are, the more trustworthy it becomes. The use of comprehensive data is essential to enhance the accuracy and efficiency of DT-based services. Taking this into account, the data standards and policies ensure that data are transferable in a seamless manner and adhere to a common format. These considerations thus affect the data life cycle and the associated data management plan.

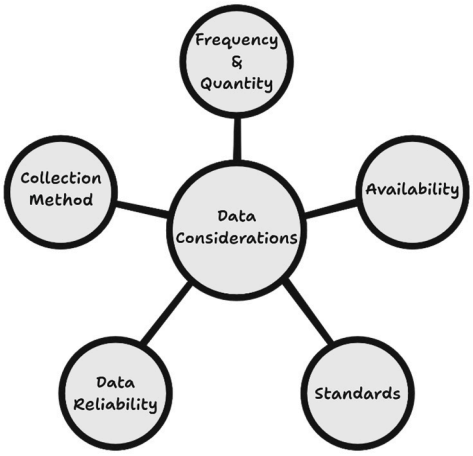

Figure 3. Important data considerations when developing a DTO.

According to the literature, data within a digital twin environment are often described by how the data are collected at the source and how often the data are sourced for input into the digital twin. In alignment with the data considerations (as shown in Figure 3) presented in the previous section, this refers to the data collection method and frequency of collection. The frequency of data collection can be grouped into three temporal categories, as follows: historical (retrospective) data, near real-time data, and real-time data [9]. The historical data category refers to data that were previously acquired or collected retrospectively. The term near real-time data refers to data that have a range between an hour to a maximum latency of a few days [22]. The term real-time data refers to data that are available for use as soon as they are generated, such as sensor observations and model predictions. The frequency of the data collection is dependent on the type of method or technologies that are employed. There are four broad categories of data collection methods, which are depicted in Figure 4. These can be described as real-time data from in situ sensors and instrumentation; modelled data; unstructured data; geospatial data; and Earth observation data [7,9,14,20]. Throughout these varied data types, each observation has a specific range and purpose and can be fused with others to provide a more comprehensive overview of the ocean's current state [17]. For this study, other geospatial data and unstructured data are grouped into ancillary data as shown in Figure 4.

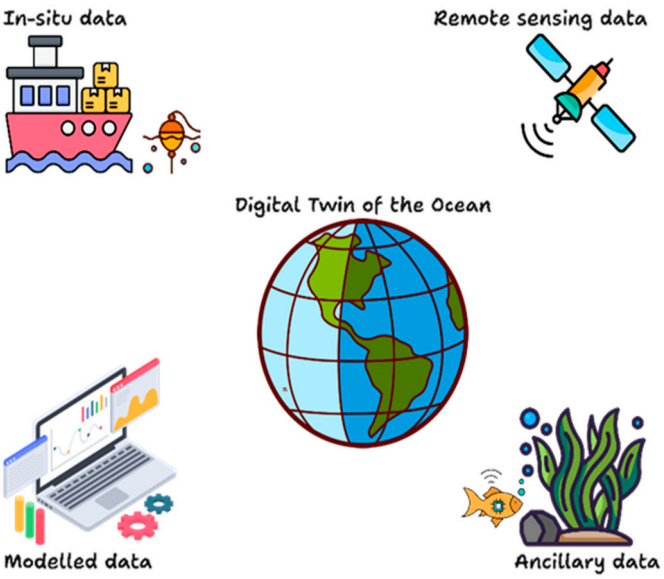

Figure 4. Various input data sources that can be used in a DTO.

The following subsections present and discuss available ocean datasets by type and perceived usage in the digital twin of the ocean. The accuracies of the various datasets are also discussed where applicable

2.1. In Situ Measurements of the Ocean

In situ measurements refer to data that is directly collected from the ocean using various types of instruments and platforms. Within the ocean domain, these data are collected using two main methods. In situ sensors can be deployed in the ocean and collect surface and sub-surface ocean measurements. Surface ocean variables are collected by instrumentation such as moorings or buoys and tidal gauges [23]. Subsurface or seafloor instrumentation such as cable observatories and autonomous underwater vehicles (AUVs) [15]. These techniques can be used to collect ocean parameters such as temperature, salinity, currents and wave heights. In addition, data collected from in situ sensors are also used as ground-truth, validation, and calibration data within remote sensing data. Examples of usage of in situ observation methods include that of Nakath et al. [24] who presented the use of dome ports to protect underwater cameras to test underwater imaging algorithms in low-ocean-depth areas. Another example is the use of autonomous Argo floats, which are used to measure ocean currents at varying ocean depths [25]. Although in situ datasets provide an accurate overview of the ocean, data acquisition is conducted manually, which only reflects features of a physical entity for a limited time and space [10]. Furthermore, the deployment and maintenance of in situ platforms can be expensive and time-consuming [1]. Measurements of these ocean parameters can be collected by in situ methods and instrumentation at various temporal and spatial scales. To overcome these limitations, remote sensing has been used to complement in situ field measurements [26].

2.2. Satellite Remote Sensing

Satellite remote sensing offers an opportunity for wide area monitoring of the oceans. There has been an increase in the amount of Earth observation satellites in the last decade that has become an important part of oceanographic monitoring [8]. The key advantages of using remote sensing datasets for the development of a DTO are the large area coverage, wide range of temporal and spatial resolutions, and the low cost of derived datasets [1].

Space-borne remote sensing data can be acquired from both passive and active instruments for several ocean applications. Ocean parameters can be collected using passive remote sensing systems, which encompass the reflected electromagnetic energy in the visible, Near-Infrared (NIR), and Shortwave Infrared (SWIR) bands, as well as emitted electromagnetic energy in the Thermal Infrared (TIR) bands. The imagery acquired can be packaged as multispectral imagery where the selection of bands is determined by the sensor radiometric resolution [1].

In the marine and oceanographic domains, passive sensors are used to derive parameters such as ocean colour, bathymetry, and sea surface topography [1]. Thermal infrared sensors can provide sea surface temperature (SST) and ocean colour satellite observations. Thermal infrared radiometer (TIR) sensors are passive remote sensing systems that can be used to provide SST observations. There are found on Low-Earth Orbiters (LEO) such as NASA's Aqua Moderate Resolution Imaging Spectroradiometer (MODIS) and the NOAA Visible Infrared Imaging Radiometer Suite (VIIRS) [25,27]. Geostationary satellite instruments have also been used to acquire operational SST estimation, such as with the Geostationary Operational Environmental Satellite (GOES) and the European Spinning Enhanced Visible and Infrared Imager (SEVIRI) [25].

More recently, ESA's Copernicus mission has provided freely available global SST and ocean colour satellite observations from Sentinel-3 Ocean Land and Colour (OLCI) and Sea

and Land Surface Temperature Radiometer (SLSTR) sensors [25]. The main limitation of passive sensors is that inclement weather and atmospheric effects hinder image quality. Additionally, passive sensors require natural energy (lighting) to acquire useful imagery.

Active sensors, such as Synthetic Aperture Radar (SAR), have a day-and-night imaging capability and operate in all weather conditions. These sensors have been used to detect a variety of features, such as oil spills, plastic pollution, and ships in the ocean [1,22,26]. SAR data have also been used to derive a variety of ocean parameters such as ocean surface wind and current, bathymetry, and ocean tidal information [28]. However, these data are challenging to interpret and could be affected by speckle noise. Speckle noise in an SAR image is multiplicative noise, and arises from the coherent sum of the ground scatterers, which are distributed randomly within each pixel [29]. This can hinder the ability to distinguish between small features and reduces the overall image quality.

Some of the limiting factors that remote sensing data presents for digital twins are the spatial and temporal resolutions. The spatial resolution is the smallest feature that can be detected by a sensor, which is represented as a pixel. The temporal resolution refers to the frequency at which a certain location is captured [8]. A limitation of using remote sensing satellite observations is that the spatial resolution is often too coarse to derive ocean parameters at a local scale [1]. Furthermore, long revisits between image acquisitions may be unsuitable for real-time or near-real-time monitoring of the ocean. Although remote sensing can provide a synoptic view, it cannot replace measurements that are directly gathered from instrumentation in the ocean.

Whilst in situ and remote sensing data provide current (real-time) and retrospective (historical) views about ocean conditions and marine objects, there is a need to predict future conditions to support maritime operations, climate change preparedness, and improve ocean health [30]. Satellite imagery plays a prominent role in machine learning research and applications. Specialised fields such as computer vision have revolutionised the applicability of remotely sensed images in object detection and classification which are used extensively in oceanic and maritime domains.

2.3. Modelled and Existing Training Data

Models can be used to aggregate the information provided by oceanic, atmospheric and terrestrial observation, propagate it through time, and extrapolate it into the unobserved variables and spatial regions [31]. In a digital twin of the ocean, various types of ocean model data are used, including numerical, empirical, statistical, physical, and machine learning model data [32–34]. These model data can be integrated to provide a comprehensive and dynamic representation of an oceanic system. Unlike conventional models, the digital twin of the ocean requires models to adapt, evolve, and learn based on new observations that are acquired [31]. Numerical models simulate ocean dynamics such as circulation, tides and waves, enabling real-time monitoring and predictive simulations essential for resource management and disaster response [35,36]. Empirical models use observational data for statistical predictions by using key oceanographic variables, while statistical models forecast long-term trends [37]. Physical models provide insights into localised processes. Machine learning, particularly deep learning, improves computational efficiency by leveraging large datasets and enabling lower numerical precision, enhancing the predictive capabilities of digital twins [38,39]. These models can be integrated to allow digital twins to perform real-time monitoring, scenario simulations, and decision support for applications such as environmental conservation and climate adaptation planning [33].

Modelled data are essential for predicting large-scale ocean events and provide a powerful tool for long-term ocean management. Global ocean models combine various data sources to simulate oceanic systems on a planetary scale. These models provide critical

information on global weather patterns, marine ecosystems, carbon cycle studies, fisheries optimisation, and global maritime safety [31,40]. Therefore, these models are vital for environmental forecasting, climate research, and sustainable ocean management. Several global models are available for ocean applications and can be integrated into a digital twin to enhance its predictive capabilities and provide a comprehensive representation of ocean dynamics [41,42]. These are the Nucleus for European Modelling of the Ocean (NEMO) [43,44], which simulates ocean dynamics and sea ice, and the Hybrid Coordinate Ocean Model (HYCOM) known for global ocean circulation and forecasting [45]. The Regional Ocean Modelling System (ROMS) [46] and Finite-Element Sea ice-Ocean Model (FESOM) [47] specialise in regional and global ocean dynamics, and Copernicus Marine Service (CMEMS) [36] offers real-time operational models, making them essential tools for monitoring, scenario testing, and long-term ocean management within a digital twin framework. With regard to machine learning models, the requirement for their use is sufficient and accurate training data. Various agencies have embarked on the collection of high-quality training data. Consequently, training datasets need to be incorporated into the digital twin as input to improve the accuracy and efficiency of machine learning models [48]. For example, global Argo data have been made freely available through the Argo programme, which has over two million vertical profiles of temperature and salinity observations from the upper two kilometres of the global ocean [49]. A benchmark dataset has also been made available for developing and evaluating machine learning algorithms capable of detecting marine debris called the Marine Debris Archive (MARIDA). This dataset, which is derived from multispectral Sentinel-2 data, includes various marine features, such as ships and waves, and has annotations from verified plastic debris events in several geographic regions [50].

The limitations of using modelled data are that any unobserved outputs provided by the model will always be more uncertain than observable inputs put into the model [31]. Furthermore, the model is often limited to a certain scale and inferred to larger regions. There are also uncertainties and biases associated with the formulation of the model. In addition to models and existing training datasets, ancillary data sources can be used to aid the development of a DTO.

2.4. Ancillary Datasets

Additional data sources provide context and enhance the utility of digital twins of the ocean. In this context, ancillary data collection can include data from organisations at sea, social media platforms and open-source databases that are publicly available. Examples of ancillary datasets are information on shipping routes, fishing activity, marine boundaries, ecological classifications. Apart from other geospatial data, ancillary data can be in an unstructured format, which may include location-based tabular data, texts, and videos describing oceanic events. Two well-known organisations that provide open-source ancillary data are the Ocean Info Hub and Marineregions.org (accessed on 30 October 2024) [51].

2.5. Summary of Various Data Sources

There are various data sources that can be used as input for the DTO, and each data type has various advantages and disadvantages that need to be considered. The objective is to use the various data sources in a complementary manner to ensure that the highest possible data quality is used. The DTO in the South African and Western Indian Ocean regions will incorporate ocean observations from both in situ and satellite platforms. The DTO in these regions is envisaged to be high-resolution, multi-dimensional, and a near real-time virtual representation of various ocean variables. In Table 1, an overview of several ocean variables, the data collection methods, and temporal and spatial

resolution are presented. Overall, the in situ methods can provide high resolution data (metres to kilometres) but data collection can often have limited spatial sampling. These measurements also only represent a portion of the ocean. Conversely, remote sensing can be acquired by both passive and active sensors, which can be used for wide area monitoring. A limitation of using passive remote sensing is that observations may be affected by cloud cover. In the instance where clouds are present, SAR satellite images could be used to acquire information about the ocean but may be difficult to interpret. Data from models and existing training datasets can be used for global modelling. However, models are only a representation of the real-world. Ancillary datasets can be used to enhance information gathered by in situ and remote sensing observations, but there is uncertainty in data quality and user bias.

Table 1. Ocean variables and the various data collection techniques, frequencies, and spatial scale/resolutions.

Ocean Variable	Data Collection Techniques	Frequency	Spatial Scale/Resolution
Sea surface temperature	In situ: surface buoy thermometer; Remote sensing: TIR and microwave radiometers	In situ: hourly/daily Remote sensing: 1 to 8 days	In situ: 1 m to 1 km Remote sensing: 750 m to 1 km
Sea salinity	In situ: conductivity temperature depth (CTD) probe/sensors	In situ: hourly/daily	In situ: 1 m to 1 km
pH	In situ: electrode	In situ: daily/weekly	In situ: depth of 2 km
Pressure	In situ: strain gauge sensor; marine pressure sensors	In situ: daily/weekly	In situ: depth of 7 km
Ocean currents and waves	In situ: acoustic Doppler current profiler (ADCP) sensors; Remote sensing: SAR	In situ hourly/daily; Remote sensing: 11 to 24 days	In situ: 1 m to several km; Remote sensing: up to 85 km
Ocean colour	Remote sensing: optical and multispectral, VIIRS, Sentinel-3 OLCI; Oceansat-2	Remote sensing: 1 to 4 days	Remote sensing: 300 m to 1 km
Wind speed and direction	Remote sensing: scatterometer; ASCAT, QuikSCAT	Remote sensing: daily to weekly	Remote sensing: 12.5 to 25 km
Wave height	Remote sensing: radar altimeter, Jason-3; Sentinel-6, CryoSat-2	Remote sensing: 10 days	Remote sensing: 10 km
Bathymetry	In situ: single/multi-beam surveying; Autonomous Underwater Vehicle (AUV); Remote sensing: airborne LiDAR, Multispectral Remote Sensing, satellite altimetry	In situ: as required; remote sensing: weekly for satellite observations; airborne LiDAR campaigns yearly	In situ: 1 to 50 m at maximum depth of 11 km; Remote sensing: 5 m to 30 m
Human activities at the ocean	GIS coverages, media and crowd-sourced data	Daily	Local to global scale datasets are available
Models	Physical and numerical models NEMO, HYCOM, ROMS, FESOM	Global scale	1/10° to 1/12° horizontal resolution

The broad nature of data that can be used in a digital twin and the variety of data sources is well encapsulated in the description of multi-source heterogeneity, which covers multi-platform, multi-parameter, multi-structure and multi-resolution characteristics [9]. This multi-heterogeneity necessitates the implementation of a data management plan that guides how data are used during the data life cycle in a digital twin environment. The data management plan is important for the successful implementation of the digital twin [52]. The following section discusses the broad aspects of data management, data management plans, and implementation of the digital twin of the ocean for the Southern African and Western Indian Ocean regions.

3. Data Management in a Digital Twin of the Ocean

In broad terms, data management refers primarily to architectures, policies, practises and approaches for managing data life cycle needs effectively [53–55]. These approaches include data reception or creation, storage, collocation, maintenance, and disposal. Effective data management is essential for executing applications that provide critical analytic information to help drive operational decision-making and strategic planning by organisations and end users. Data management has also grown in significance, as private and government organisations are obligated to comply with regulatory requirements that include privacy and protection laws such as the Protection of Personal Information Act, in the case of South Africa, and the General Data Protection Regulation in the European Union [56].

The rapid growth and adoption of data management across various fields has introduced variability in how data management policies are adopted and implemented [9,57]. This is due to the collaborative nature of these policies, and the need to maintain integrity and uniformity. Hence, it is essential to establish best practises to work efficiently and adhere to industry standards.

3.1. Current Data Management Practises for Digital Twins

The data management component of a digital twin narrows the gap between the physical system, the mirrored digital one, and the services component [58]. With the multi-heterogeneity of data in a DTO, data management plays an integral part in the maintenance of data and processing the data into useful information. A data management plan in a digital twin includes the data collection and acquisition, the data storage and infrastructure, and the data integration techniques, as well as making sure that data governance principles are adhered to. Efficient data management does not only involve data storage, organisation, and processing, but also includes efficient access and retrieval of the information to the users [19]. The data management functionalities are either explicitly stated in a dedicated management component or included in other components of the DT [18,58].

3.1.1. Data Collection and Acquisition

The data collection and acquisition component of the data management strategy identifies the various data sources and defines the different types of data formats to account for the heterogeneity of data [58]. The data acquisition component also accounts for the various methods of collecting data from physical environments, temporal aspects, and accuracies thereof [20].

3.1.2. Data Storage

Data storage requirements depend on the digital twin's functionality [59]. This is often referred to as the data pool and serves as a centralised storage repository that consolidates data from the various sources into a standardised format [9]. This data pool makes use of

data descriptions that are based on data collection and acquisition information. Furthermore, the necessary constraints relating to the data characteristics should be outlined, in order to identify the conditions under which the data are not useful [10]. The management and organisation of data are made more efficient by establishing a central data repository. However, integrating these diverse data types into a digital twin of the ocean involves sophisticated data fusion techniques, real-time data assimilation, and advanced computational methods [60]. Cloud services have provided an advantage in the way data can be collected, stored, and managed. Large volumes of data can be stored and easily accessed in a centralised or distributed location, which improves efficiency. Users can store large volumes of data without hardware limitations, and can access data easily, provided that there is an internet connection. Cloud platforms also offer scalability and flexibility to support the digital twin as it evolves [20].

3.1.3. Data Integration

Data integration can be defined as the combination of data from multiple sources to provide a high-level unified view [58]. This process can be performed through data fusion, which merges data sources with various relations. For example, readings from in situ instrumentation can be fused with remote sensing observations. Data fusion techniques can be used to detect any anomalies, transform data into a standard format, perform data matching through metadata models such as schemas and ontologies, and carry out semantic data enrichment [10,19,58]. Existing work on data fusion approaches include methods such as the weighted average method, Bayesian methods, and neural networks [20]. For satellite based observations, several feature extraction techniques have been developed that can be integrated using mathematical and machine-learning data fusion techniques [4,27].

3.1.4. Data Access and Retrieval

The data access and retrieval function in the digital twin provides access to data ingested in the twin, to the related services layer, or directly to the end user. This is often facilitated by application programming interfaces linked to the diverse data products within the twin and are connected to the application layer where user interaction happens. Data access is often provided in real-time, automatically, or through a user-initiated request [8,9].

3.2. Data Management in the Digital Twin of the Southern African and Western Indian Ocean Region

The data management plan adopted for the digital twin of the ocean for the Southern African and Western Indian Ocean regions is aligned to a modified geospatial data life cycle, depicted in Figure 5. This data life cycle describes the process that data encounters throughout the digital twin.

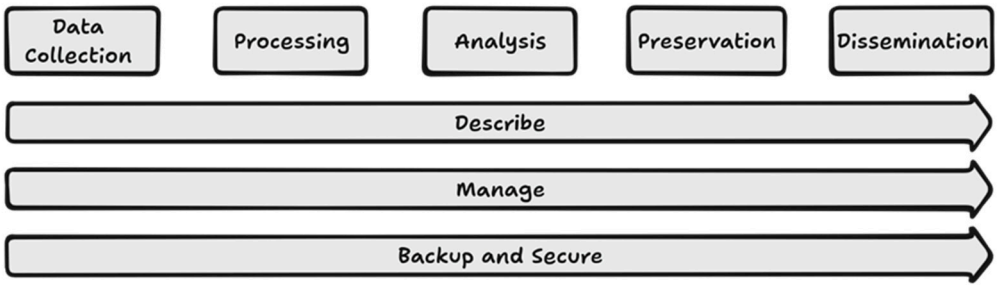

Figure 5. Data life cycle in the digital twin of Southern African and Western Indian Ocean region.

3.3. Data Life Cycle

The first step of the cycle is data collection, which describes the acquisition of data from various data sources. Access to external data sources is primarily gained through web-based Application Programming Interfaces (APIs) with an authentication layer, for proprietary data. Open portals often require initial registration for user management. After data are ingested, they are processed to ensure that they are suitable for assimilation and compatible with the digital twin environments. In addition, data should adhere to the prescriptions stipulated by the data description, management, backup, and secure undercurrents. Following data collection is data processing, which is generally automated to ensure seamless use in subsequent steps of the life cycle. The third step of the data life cycle is data analysis, where data are analysed to produce value-added products that are generated by different models, as described in Section 2.3 of this paper. The fourth step describes how data are stored in the digital twin by considering their heterogenous nature. The fifth step of the data life cycle focuses on dissemination of data and value-added products. The dissemination includes making the pre-processed data available where permissible, based on data licencing and agreements. Dissemination of data and products is enabled by using internally developed APIs, which form part of the digital twin. It is worth noting that these processes are not linear. For example, data that does not require processing can be ingested into the data storage directly, bypassing the processing step. Some models require retrospective time series data and, thus, access data from storage, whilst others process real time data streams. These processed data, along with the analysed results, are stored afterwards and then disseminated when required. Throughout the life cycle, there are undercurrents that underpin all activities; these are description of data, data management, and backup and security. Description of data occurs every time data change states, for instance, on initial ingestion, after pre-processing, or when new products are generated from analysis. This is performed to preserve the provenance of data throughout the digital twin. Another undercurrent is data management, which ensures data accessibility, reliability, and adherence to standards and specifications throughout the life cycle to ensure high data quality and interoperability. The final undercurrent, backup and security, aims to protect the data against losses and breaches.

4. Data Management Plan

Given the data life cycle described above, the data management plan encompasses the different phases that guide actions and operations for effective management of the data throughout this cycle. The data management plan steps are discussed henceforth and are aligned with Michener's [61] ten rules for creating a good data management plan and the Group on Earth Observations' data management principles [62]. This data management plan is illustrated in this section.

4.1. Principle 1: Conceptualisation

The first step before the collection of data is to identify essential data, based on the requirements for the regional digital twin of the ocean. These include internally generated products and the trained models. These data and products are largely informed by stakeholder needs and key activities in the region. These include fisheries and aquaculture support, marine and maritime safety and security, and pollution, water quality, and other coastal monitoring activities. The listing of important services is derived from interactions with users of the South African Oceans and Coasts Information System (OCIMS) [63], Global Monitoring of the Environment and Security (GMES), and Africa Marine and Coastal Operations for Southern African and Western Indian Ocean (MarCOSIO).

4.2. Principle 2: Identify Data to Be Collected

This task includes the listing of the various data requirements captured in principle 1. The list provides information on the main sources of the data, alternative sources of the data, and the formats in which the data are sourced. Additional information captured about the data includes the volume of data from the source and the frequency of reception. Additionally, APIs and other methods of data access are documented, which aligns with the data collection step of the data life cycle.

4.3. Principle 3: Organisation of Data

The organisation of data is planned for and documented once there is a comprehensive understanding of the characteristics of the data to be collected. As a result of the multi-heterogenous nature of the data collected, several data storage mechanisms are considered for the digital twin of the ocean. Most of the data are stored in multidimensional data cubes to account for the time series and depth measurements that are found in most ocean datasets. Relational databases are also used to store maritime datasets, such as vessel movement information. The idea is to store the data close to its original form following the extract, load, and transform model (ELT). This is achieved whilst minimising the amount of initial processing and preserving it for later use by the models. As described in the data description section, marine and maritime datasets are stored in well-known standardised formats and require very little transformation before they are stored. This task forms part of the processing and preservation of data in the data life cycle.

4.4. Principle 4: Documentation of Data Descriptions

Thoroughly describing data ensures that it can be easily discovered and retrieved. The process can be undertaken concurrent to the organisation of the data in suitable databases for preservation purposes; in this case, PostgreSQL with GIS extensions (PostGIS) is preferred. This DTO makes use of standards-compliant metadata formats to describe the data. Metadata capturing is well-documented and described in geospatial data infrastructures. Some of the data used in DTOs are acquired from non-geospatial sources and thus have no spatial descriptions. Therefore, considerations are made for metadata descriptions from relevant related domains such as oceanography and meteorology, such as the climate and forecasting (CF) metadata conventions.

4.5. Principle 5: Data Quality Assurance

It is essential to describe the quality of the data and the products generated within the digital twin environment. According to the GEO data management principles for Earth observation data, quality control should verify consistency, accuracy, completeness, correctness, and the fitness of use of a dataset. This view is largely supported by the ISO 19157 [63] data quality standard and the ISO 19158 [64] quality assurance of data supply standards for geospatial data. The ISO 19157 data quality standard defines the five elements of quality as completeness, logical consistency, positional accuracy, thematic accuracy, and temporal accuracy. ISO 19158 describes a framework for describing the quality assessment process that a data supplier follows and communicates to the data consumer, in the data production process [65,66].

Data acquired from meteorological and oceanographic sources often contain a statement on some of these elements. However, quality control is still largely dependent on the data providers and the metadata received about a dataset. Model accuracy is assumed to be as good as the input data sources used.

4.6. Principle 6: Data Storage and Preservation of Statergy

Related to the organisation of the data (principle 3) is the decision and strategy for preservation of the data. This describes the continuity plan for the storage and accessibility of data in the digital twin and the long-term preservation plan. It answers questions about how long the data will be accessible, how the data will be stored and protected for the accessibility period, and preservation thereafter. As a result of the relationship between the OCIMS and MarCOSIO projects, this DTO is expected to have long-term sustainability due to the support of these long-term programmes. Preservation of data in the DTO is expected to be managed through the provisioning of additional data servers, with one dedicated to archiving older products with low accessibility demands. This task is relevant to the data preservation step of the data life cycle.

4.7. Principle 7: Data Policies and Governance

Data policies govern how data will be managed and shared. These are stipulated by the data providers and include licencing arrangements, plans for sharing and retaining the data, and legal and ethical use restrictions. The policies are set out in the data agreements that are issued when access is granted. This DTO mainly makes use of open data available under Open Data Commons licences. However, proprietary data are also ingested in accordance with user needs for very-high-resolution information. Policies on ethical usage and data mandates are essential and cannot be ignored. This DTO therefore takes into consideration regional and local data mandates and policies on provision and dissemination of data, to ensure ethical use. These policies mostly apply to and influence the analysis, preservation and dissemination steps of the data life cycle.

4.8. Principle 8: Description of Data Dissemination

Dissemination of data and information products is the most public facing activity in the DTO driving usability and user satisfaction. According to Michener [61], a good dissemination plan should state availability and accessibility restrictions, but should be as minimal as possible. This answers questions relating to what data are available and the frequency of availability. This DTO of the Southern African and Western Indian Ocean regions uses the FAIR and TRUST principles to ensure sustainable data discovery and accessibility. Data from various sources are adequately described, and standardized with regard to compliancy before they are disseminated. This DTO follows the principles outlined in Sibolla et al. [67] to ensure alignment and conformity. FAIR implementation is guided by the criteria of Jacobsen et al. [68] and facilitated through the Comprehensive Knowledge Archive Network (CKAN). Periodic updates and synchronisation of the CKAN records is performed through an ISO metadata-harvesting process to ensure that only the latest descriptions are disseminated. Similarly, TRUST enforcement is guided by Li et al.'s work [69], and ensures that the DTO's data are reliable and trustworthy. This is to ensure that this DTO receives and disseminates accurate and reliable data.

4.9. Principle 9: Management of Roles and Responbsibilities

This principle maps out the roles that different members of the technical team, stakeholders, and users of the DTO assume. The roles may map to the functions required for the data life cycle and include data collection, data entry, quality assurance, metadata creation and management, backup, data preparation and submission to an archive, and systems administration.

The mapping of the data management principles and tasks to the data life cycle is described in Figure 6. According to the discussion above, it is evident that the data

management activity underpins the data life cycle in a DTO and shows how this relates to the various data management principles discussed.

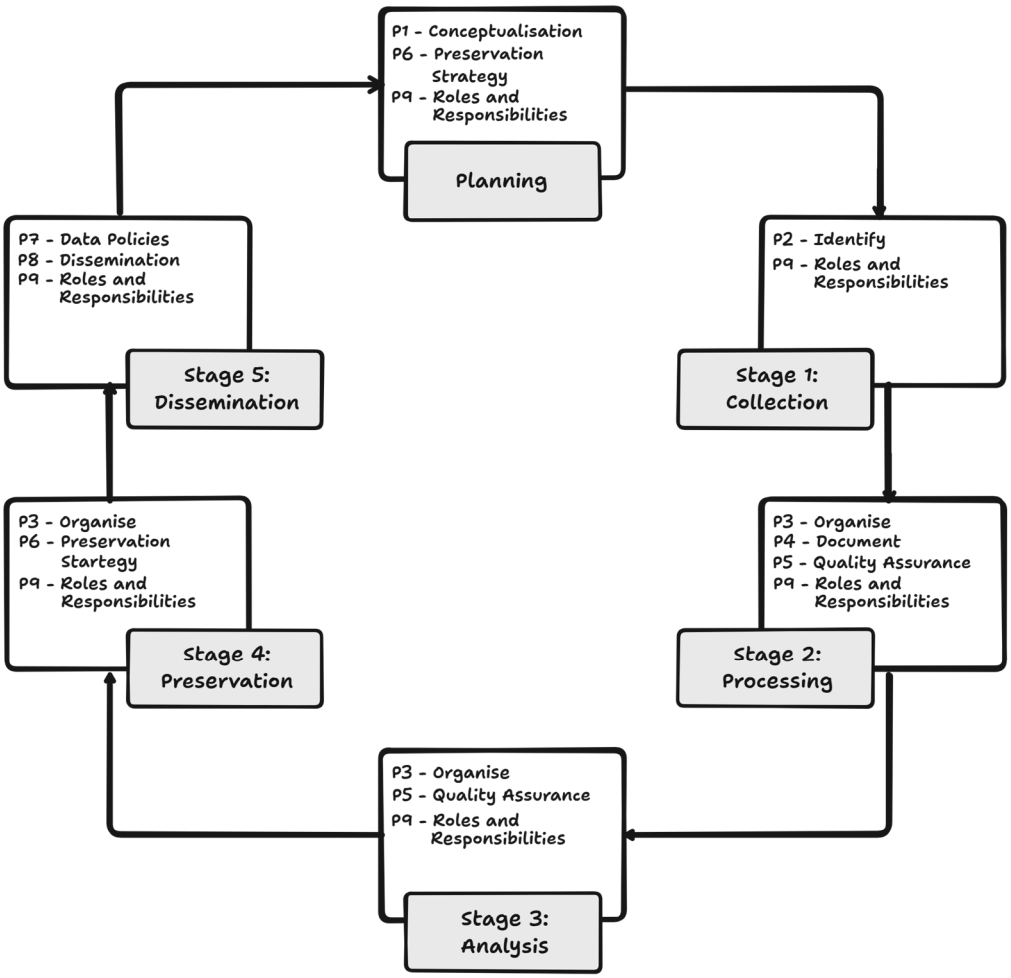

Figure 6. Mapping of the data management principles and tasks in relation to the data life cycle.

5. Challenges of Data Management in a Digital Twin of the Ocean

This section describes the challenges of data, and the management thereof, in a digital twin of the ocean.

Firstly, different modes of data collection, differing imaging resolutions, and differing temporal resolutions pose challenges with regard to data fusion. The process of harmonising this type of data often requires transformation and alignment, which is often accompanied by data loss and misalignments. Temporal resolution differences often require some interpolation methods to account for data gaps, which reduces the overall accuracy in interpolated regions. In addition to temporal variations, radiometric resolution poses a challenge, specifically when used as inputs in models that operate on point data. Remotely sensed data are generally affected by pixel mixing at lower resolutions. Consequently, these data do not model point measurements accurately.

The various formats in which the data are presented also pose a challenge with regard to storage. This often requires an additional processing step of harmonising the data formats.

The distributed responsibility of ensuring data quality also poses a challenge. While it is near impossible to have centralised data quality assurance, having distributed data quality assurance does not ensure high quality. Full TRUST compliance, as discussed in Sibolla et al. [70], guarantees high levels of quality assurance and long-term sustainability.

Another challenge relates to referencing non-spatial information sourced from non-geospatial entities. Often, useful information is gathered through various channels, some of which do not contain spatial information. However, the location of this information is vital for fusion with other spatially referenced sources.

Finally, while machine learning models have provided feature rich networks for learning various insights about the data, they often rely on the collection of reliable training data on areas of interest. Limited training data for some areas of interest thus limits the efficiency of some trained models. However, there are collaborative drives that are aimed at sourcing and sharing training data to ensure accurate results.

In summary, the challenges regarding data in a DTO are as follows: ocean data are very complex to standardise; this is in agreement with the findings of Chen et al. [9]. Furthermore, inconsistency and uncertainty related to data need to be reviewed and accounted for; hence, there is a strong need for quality control and quality assurance [10]. Data integration poses a challenge, because data can be acquired from various domains, therefore highlighting the need for interoperability measures to be put in place. Data quality assurance may be the sole responsibility of the provider in cases where metadata records are harvested and shared.

6. Conclusions

This paper presents the first step towards the development of a DTO for Southern African and the Western Indian Ocean regions. This will enable wide area monitoring of the ocean and aid in decision making. Data are central to the successful implementation and usability of the DTO. Hence, as a first step, it is important to understand the requirements, characteristics, and domain considerations of the DTO regarding how the data will be used. For this purpose, a data landscape study was undertaken, which revealed the multi-heterogeneity of the data required for this DTO. Unpacking the multi-heterogeneity aspects revealed that the data required for this digital twin are not only multi-disciplinary in origin, but the data are also multi-source, multi-structured, multi-platform, multi-resolution, multi-parametric, and multi-dimensional in nature. As a result of these characteristics, a suitable data management plan needed to be developed and implemented, in order to support the phases of data usage in the DTO. The phases of the data usage in the DTO are encapsulated in the data life cycle, which includes data collection, processing, analysis, preservation, and dissemination. The resultant data management plan provides an undercurrent to this life cycle by outlining the relevant data management principles. The main considerations in the nine-point data management plan include mechanisms for describing data through standardised metadata, data quality assurance, data organisation and storage, and roles and responsibilities for operational purposes. The development of the data management plan revealed limitations in data quality assurance, which is largely assumed to be managed at the source; however, mitigation mechanisms involve making provisions for testing. In conclusion, the requirement for standardisation improves interoperability, and extends usability. Furthermore, the DTO domain requires understanding of the standard landscapes across several data-producing domains. This data management plan is based on the scientific literature and published operational practises within the Earth observation science domain, and it addresses the DTO domain data characteristics, operation environment, and user expectations.

Author Contributions: Conceptualization, Shelley Haupt, Bolelang Sibolla, and Raymond Nkadimeng Molapo; methodology, Bolelang Sibolla, Raymond Nkadimeng Molapo, and Shelley Haupt; formal analysis, Shelley Haupt, Bolelang Sibolla, and Raymond Nkadimeng Molapo; investigation, Shelley Haupt, Bolelang Sibolla, Raymond Nkadimeng Molapo, Lizwe Mdakane, and Nicolene Fourie; resources, Shelley Haupt, Bolelang Sibolla, Raymond Nkadimeng Molapo, Lizwe Mdakane, and

Nicolene Fourie; data curation, Shelley Haupt, Bolelang Sibolla, Raymond Nkadimeng Molapo, Lizwe Mdakane, and Nicolene Fourie; writing—original draft preparation, Shelley Haupt, Bolelang Sibolla, Raymond Nkadimeng Molapo, Lizwe Mdakane, and Nicolene Fourie; writing—review and editing, Shelley Haupt, Bolelang Sibolla, and Raymond Nkadimeng Molapo; supervision, Bolelang Sibolla and Nkadimeng Raymond Molapo. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The original contributions presented in the study are included in the article. Further inquires can be directed to the corresponding authors.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Amani, M.; Moghimi, A.; Mirmazloumi, S.M.; Ranjgar, B.; Ghorbanian, A.; Ojaghi, S.; Ebrahimy, H.; Naboureh, A.; Nazari, M.E.; Mahdavi, S.; et al. Ocean Remote Sensing Techniques and Applications: A Review (Part I). *Water* **2022**, *14*, 3400. [CrossRef]
2. Costanza, R. The ecological, economic and social importance of the oceans. *Ecol. Econ.* **1999**, *31*, 199–213.
3. Tzachor, A.; Hendel, O.; Richards, C.E. Digital twins: A stepping stone to achieve ocean sustainability? *npj Ocean Sustain.* **2023**, *2*, 16. [CrossRef]
4. Lv, Z.; Lv, H.; Fridenfalk, M. Digital Twins in the Marine Industry. *Electronics* **2023**, *12*, 2025. [CrossRef]
5. Teh, L.C.L.; Sumaila, U.R. Contribution of marine fisheries to worldwide employment. *Fish Fish.* **2013**, *14*, 77–88. [CrossRef]
6. Ritchie, H.; Samborska, V.; Roser, M. Plastic Pollution. Available online: https://ourworldindata.org/plastic-pollution (accessed on 3 November 2024).
7. Mills, G.; Fones, G. A review of in situ/IT methods and sensors for monitoring the marine environment. *Sens. Rev.* **2012**, *32*, 17–28. [CrossRef]
8. Miedtank, A.; Schneider, J.; Manss, C.; Zielinski, O. Marine digital twins for enhanced ocean understanding. *Remote Sens. Appl. Soc. Environ.* **2024**, *36*, 101268. [CrossRef]
9. Chen, G.; Yang, J.; Huang, B.; Ma, C.; Tian, F.; Ge, L.; Xia, L.; Li, J. Toward digital twin of the ocean: From digitalization to cloning. *Intell. Mar. Technol. Syst.* **2023**, *1*, 3. [CrossRef]
10. Zhang, M.; Tao, F.; Huang, B.; Liu, A.; Wang, L.; Anwer, N.; Nee, A.Y.C. Digital twin data: Methods and key technologies. *Digit. Twin* **2022**, *1*, 2. [CrossRef]
11. Liu, M.; Fang, S.; Dong, H.; Xu, C. Review of digital twin about concepts, technologies, and industrial applications. *J. Manuf. Syst.* **2021**, *58*, 346–361. [CrossRef]
12. Grieves, M. Digital Twin: Manufacturing Excellence through Virtual Factory Replication. *White Paper* **2014**, *1*, 1–7.
13. Bahurel, P.; Brönner, U.; Buttigieg, P.-L.; Chai, F.; Chassignet, E.; Devey, C.; Fanjul, E.A.; Hill, K.; Kim, S.Y.; Kollert, J.; et al. DITTO Programme White Paper. Available online: https://ditto-oceandecade.org/wp-content/uploads/2023/07/DITTO-Whitepaper_FINAL-July-2023.pdf (accessed on 3 October 2024).
14. Berre, A.J.; Pearlman, J.; Bye, B.; Masó, J.; Digital, S. Iliad Digital Twins of the Ocean Interoperability Architecture. In Proceedings of the IGARSS, Athens, Greece, 7–12 July 2024; pp. 3568–3571.
15. Yu, Z.; Du, P.; Yi, L.; Luo, W.; Li, D.; Zhao, B.; Li, L.; Zhang, Z.; Zhang, J.; Zhang, J.; et al. Coastal Zone Information Model: A comprehensive architecture for coastal digital twin by integrating data, models, and knowledge. *Fundam. Res.* **2024**. [CrossRef]
16. Spanoudaki, K.; Kozyrakis, G.; Metheniti, V.; Parasyris, A.; Kampanis, N. The Cretan Sea oil spill Digital Twin pilot for the ILIAD Digital Twin of the Ocean. In Proceedings of the Cretan Sea Oil Spill Digital Twin Pilot for the ILIAD Digital Twin of the Ocean, EGU General Assembly, Vienna, Austria, 23–28 April 2023.
17. Brönner, U.; Sonnewald, M.; Visbeck, M. Digital Twins of the Ocean can foster a sustainable blue economy in a protected marine environment. *Int. Hydrogr. Rev.* **2023**, *29*, 26–40. [CrossRef]
18. Tao, F.; Zhang, M. Digital Twin Shop-Floor: A New Shop-Floor Paradigm Towards Smart Manufacturing. *IEEE Access* **2017**, *5*, 10.
19. Singh, S.; Shehab, E.; Higgins, N.; Fowler, K.; Reynolds, D.; Erkoyuncu, J.A.; Gadd, P. Data management for developing digital twin ontology model. *Proc. Inst. Mech. Eng. Part B J. Eng. Manuf.* **2021**, *235*, 2323–2337. [CrossRef]
20. Dihan, M.S.; Akash, A.I.; Tasneem, Z.; Das, P.; Das, S.K.; Islam, M.R.; Islam, M.M.; Badal, F.R.; Ali, M.F.; Ahamed, M.H.; et al. Digital twin: Data exploration, architecture, implementation and future. *Heliyon* **2024**, *10*, e26503. [CrossRef]
21. Liu, Y.; Qiu, M.; Liu, C.; Guo, Z. Big data challenges in ocean observation: A survey. *Pers. Ubiquitous Comput.* **2017**, *21*, 55–65. [CrossRef]
22. Wu, X.; Lu, G.; Wu, Z. Remote Sensing Technology in the Construction of Digital Twin Basins: Applications and Prospects. *Water* **2023**, *15*, 2040. [CrossRef]

23. Lin, M.; Yang, C. Ocean Observation Technologies: A Review. *Chin. J. Mech. Eng.* **2020**, *33*, 32. [CrossRef]
24. Nakath, D.; She, M.; Song, Y.; Köser, K. An Optical Digital Twin for Underwater Photogrammetry: GEODT—A Geometrically Verified Optical Digital Twin for Development, Evaluation, Training, Testing and Tuning of Multi-Media Refractive Algorithms. *PFG—J. Photogramm. Remote Sens. Geoinf. Sci.* **2022**, *90*, 69–81. [CrossRef]
25. Minnett, P.J.; Alvera-Azcárate, A.; Chin, T.M.; Corlett, G.K.; Gentemann, C.L.; Karagali, I.; Li, X.; Marsouin, A.; Marullo, S.; Maturi, E.; et al. Half a century of satellite remote sensing of sea-surface temperature. *Remote Sens. Environ.* **2019**, *233*, 111366. [CrossRef]
26. Biermann, L.; Clewley, D.; Martinez-Vicente, V.; Topouzelis, K. Finding Plastic Patches in Coastal Waters using Optical Satellite Data. *Sci. Rep.* **2020**, *10*, 5364. [CrossRef]
27. Amani, M.; Mehravar, S.; Asiyabi, R.M.; Moghimi, A.; Ghorbanian, A.; Ahmadi, S.A.; Ebrahimy, H.; Moghaddam, S.H.A.; Naboureh, A.; Ranjgar, B.; et al. Ocean Remote Sensing Techniques and Applications: A Review (Part II). *Water* **2022**, *14*, 3401. [CrossRef]
28. Asiyabi, R.M.; Ghorbanian, A.; Tameh, S.N.; Amani, M.; Jin, S.; Mohammadzadeh, A. Synthetic Aperture Radar (SAR) for Ocean: A Review. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 9106–9138. [CrossRef]
29. Saxena, N.; Rathore, N. A review on speckle noise filtering techniques for SAR Images. *Int. J. Adv. Res. Comput. Sci. Electron. Eng.* **2013**, *2*, 243–247.
30. Link, J.S.; Thur, S.; Matlock, G.; Grasso, M. Why we need weather forecast analogues for marine ecosystems. *ICES J. Mar. Sci.* **2023**, *80*, 2087–2098. [CrossRef]
31. Skákala, J.; Awty-Carroll, K.; Menon, P.P.; Wang, K.; Lessin, G. Future digital twins: Emulating a highly complex marine biogeochemical model with machine learning to predict hypoxia. *Front. Mar. Sci.* **2023**, *10*, 1058837. [CrossRef]
32. Ko, D.S.; Martin, P.J.; Rowley, C.D.; Preller, R.H. A real-time coastal ocean prediction experiment for MREA04. *J. Mar. Syst.* **2008**, *69*, 17–28. [CrossRef]
33. Sonnewald, M.; Lguensat, R.; Jones, D.C.; Dueben, P.D.; Brajard, J.; Balaji, V. Bridging observations, theory and numerical simulation of the ocean using machine learning. *Environ. Res. Lett.* **2021**, *16*, 073008. [CrossRef]
34. Burchard, H.; Rennau, H. Comparative quantification of physically and numerically induced mixing in ocean models. *Ocean Model.* **2008**, *20*, 293–311. [CrossRef]
35. Armenio, E.; Ben Meftah, M.; De Padova, D.; De Serio, F.; Mossa, M. Monitoring systems and numerical models to study coastal sites. *Sensors* **2019**, *19*, 1552. [CrossRef]
36. Lellouche, J.M.; Greiner, E.; Le Galloudec, O.; Garric, G.; Regnier, C.; Drevillon, M.; Benkiran, M.; Testut, C.E.; Bourdalle-Badie, R.; Gasparin, F.; et al. Recent updates to the Copernicus Marine Service global ocean monitoring and forecasting real-time 1g 12° high-resolution system. *Ocean Sci.* **2018**, *14*, 1093–1126. [CrossRef]
37. Jonathan, P.; Ewans, K. Statistical modelling of extreme ocean environments for marine design: A review. *Ocean Eng.* **2013**, *62*, 91–109. [CrossRef]
38. Li, X.; Liu, H.; Wang, W.; Zheng, Y.; Lv, H.; Lv, Z. Big data analysis of the Internet of Things in the digital twins of smart city based on deep learning. *Future Gener. Comput. Syst.* **2022**, *128*, 167–177. [CrossRef]
39. Moore, A.M.; Martin, M.J.; Akella, S.; Arango, H.G.; Balmaseda, M.; Bertino, L.; Ciavatta, S.; Cornuelle, B.; Cummings, J.; Frolov, S.; et al. Synthesis of ocean observations using data assimilation for operational, real-time and reanalysis systems: A more complete picture of the state of the ocean. *Front. Mar. Sci.* **2019**, *6*, 90. [CrossRef]
40. Capotondi, A.; Jacox, M.; Bowler, C.; Kavanaugh, M.; Lehodey, P.; Barrie, D.; Brodie, S.; Chaffron, S.; Cheng, W.; Dias, D.F.; et al. Observational Needs Supporting Marine Ecosystems Modeling and Forecasting: From the Global Ocean to Regional and Coastal Systems. *Front. Mar. Sci.* **2019**, *6*, 623. [CrossRef]
41. Abraham, J.P.; Baringer, M.; Bindoff, N.L.; Boyer, T.; Cheng, L.J.; Church, J.A.; Conroy, J.L.; Domingues, C.M.; Fasullo, J.T.; Gilson, J.; et al. A review of global ocean temperature observations: Implications for ocean heat content estimates and climate change. *Rev. Geophys.* **2013**, *51*, 450–483. [CrossRef]
42. Arbic, B.K. Incorporating tides and interval gravity waves within global ocean general circulation models: A review. *Prog. Oceanogr.* **2022**, *206*, 102824. [CrossRef]
43. Madec, G.; The NEMO Team. NEMO Ocean Engine Reference Manual. *Zenodo* **2024**. [CrossRef]
44. Bessières, L.; Leroux, S.; Brankart, J.M.; Molines, J.M.; Moine, M.P.; Bouttier, P.A.; Penduff, T.; Terray, L.; Barnier, B.; Sérazin, G. Development of a probabilistic ocean modelling system based on NEMO 3.5: Application at eddying resolution. *Geosci. Model Dev.* **2017**, *10*, 1091–1106. [CrossRef]
45. Chassignet, E.P.; Hurlburt, H.E.; Metzger, E.J.; Smedstad, O.M.; Cummings, J.A.; Halliwell, G.R.; Bleck, R.; Baraille, R.; Wallcraft, A.J.; Lozano, C.; et al. Global ocean prediction with the hybrid Coordinate Ocean Model (HYCOM). *Oceanography* **2009**, *22*, 64–75. [CrossRef]
46. Haidvogel, D.B.; Arango, H.; Budgell, W.P.; Cornuelle, B.D.; Curchitser, E.; Di Lorenzo, E.; Fennel, K.; Geyer, W.R.; Hermann, A.J.; Lanerolle, L.; et al. Ocean forecasting in terrain-following coordinates: Formulation and skill assessment of the Regional Ocean Modeling System. *J. Comput. Phys.* **2008**, *227*, 3595–3624. [CrossRef]

47. Danilov, S.; Sidorenko, D.; Wang, Q.; Jung, T. The finite-volume sea ice-Ocean model (FESOM2). *Geosci. Model Dev.* **2017**, *10*, 765–789. [CrossRef]
48. van Dinter, R.; Tekinerdogan, B.; Catal, C. Predictive maintenance using digital twins: A systematic literature review. *Inf. Softw. Technol.* **2022**, *151*, 107008. [CrossRef]
49. Wong, A.P.S.; Wijffels, S.E.; Riser, S.C.; Pouliquen, S.; Hosoda, S.; Roemmich, D.; Gilson, J.; Johnson, G.C.; Martini, K.; Murphy, D.J.; et al. Argo Data 1999–2019: Two Million Temperature-Salinity Profiles and Subsurface Velocity Observations From a Global Array of Profiling Floats. *Front. Mar. Sci.* **2020**, *7*, 700. [CrossRef]
50. Kikaki, K.; Kakogeorgiou, I.; Mikeli, P.; Raitsos, D.E.; Karantzalos, K. MARIDA: A benchmark for Marine Debris detection from Sentinel-2 remote sensing data. *PLoS ONE* **2022**, *17*, e0262247. [CrossRef]
51. Flanders Marine Institute: MarineRegions.org. Available online: www.marineregions.org (accessed on 7 October 2024).
52. Huang, S.; Wang, G.; Yan, Y.; Fang, X. Blockchain-based data management for digital twin of product. *J. Manuf. Syst.* **2020**, *54*, 361–371. [CrossRef]
53. Diène, B.; Rodrigues, J.J.P.C.; Diallo, O.; Ndoye, E.H.M.; Korotaev, V.V. Data management techniques for Internet of Things. *Mech. Syst. Signal Process.* **2020**, *138*, 106564. [CrossRef]
54. Corrall, S. Designing libraries for research collaboration in the network world: An exploratory study. *Lib. Q.* **2014**, *24*, 17–48. [CrossRef]
55. Cox, A.M.; Tam, W.W.T. A critical analysis of lifecycle models of the research process and research data management. *Aslib J. Inf. Manag.* **2018**, *70*, 142–157. [CrossRef]
56. Buys, C.M.; Shaw, P.L. Data Management Practices Across an Institution: Survey and Report. *J. Librariansh. Sch. Commun.* **2015**, *3*, 1225. [CrossRef]
57. Reichmann, S.; Klebel, T.; Hasani-Mavriqi, I.; Ross-Hellauer, T. Between administration and research: Understanding data management practices in an institutional context. *J. Assoc. Inf. Sci. Technol.* **2021**, *72*, 1415–1431. [CrossRef]
58. Correia, J.B.; Abel, M.; Becker, K. Data management in digital twins: A systematic literature review. *Knowl. Inf. Syst.* **2023**, *65*, 3165–3196. [CrossRef]
59. Boyes, H.; Watson, T. Digital twins: An analysis framework and open issues. *Comput. Ind.* **2022**, *143*, 103763. [CrossRef]
60. Grossmann, V.; Nakath, D.; Urlaub, M.; Oppelt, N.; Koch, R.; Köser, K. Digital twinning in the ocean-challenges in multimodal sensing and multiscale fusion based on faithful visual models. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2022**, *5*, 345–352. [CrossRef]
61. Michener, W.K. Ten Simple Rules for Creating a Good Data Management Plan. *PLoS Comput. Biol.* **2015**, *11*, e1004525. [CrossRef]
62. Group on Earth Observations (GEO). Revised GEO Data Sharing and Data Management Principles. In Proceedings of the 23rd Programme Board Meeting, Virtual, 21–23 June 2022.
63. *ISO 19157:2023*; Geographic Information—Data Quality. ISO: Geneva, Switzerland, 2023. Available online: https://www.iso.org/standard/78900.html (accessed on 19 March 2025).
64. *ISO/TS 19158:2012*; Geographic Information—Quality Assurace of Data Supply. ISO: Geneva, Switzerland, 2012. Available online: https://www.iso.org/standard/32576.html (accessed on 19 March 2025).
65. Krug, M.; Naidoo, A.; Williams, L. South Africa's oceans and coastal and information management system towards improved ocean access, protection, and governance. *J. Environ. Manag.* **2024**, *354*, 120255. [CrossRef]
66. Lee, Y.; Choi, M.H.; Song, Y.S.; Lee, J.G.; Park, J.Y.; Li, K.J. Building an Indoor Digital Twin—A Use-Case for a Hospital Digital Twin to Analyze COVID-19 Transmission. *ISPRS Int. J. Geo-Inf.* **2024**, *13*, 460. [CrossRef]
67. Coetzee, S.; Cooper, A.K.; Rautenbach, V. Part C: Standards for Fundamental Geo-Spatial Datasets. In *Guidelines of Best Practice for the Acquisition, Storage, Maintenance and Dissemenation of Fundamental Geo-Spatial Datasets*; Clarke, D.G., Ed.; Mapping Africa for Africa (MAfA); United Nations Economic Commission for Africa (UN ECA): Addis Ababa, Ethiopia, 2014; p. 124.
68. Jacobsen, A.; de Miranda Azevedo, R.; Juty, N.; Batista, D.; Coles, S.; Cornet, R.; Courtot, M.; Crosas, M.; Dumontier, M.; Schultes, E.; et al. FAIR Principles: Interpretations and Implementation Considerations. *Data Intell.* **2020**, *2*, 10–29. [CrossRef]
69. Lin, D.; Crabtree, J.; Dillo, I.; Downs, R.R.; Edmunds, R.; Giaretta, D.; De Giusti, M.; L'Hours, H.; Hugo, W.; Jenkyns, R.; et al. The TRUST Principles for digital repositories. *Sci. Data* **2020**, *7*, 144. [CrossRef]
70. Sibolla, B.; Molapo, R.; Vhengani, L.; Mdakane, L. Systems and Architectural Support for Open Data Principles: A Marine Earth Observation Perspective. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.—ISPRS Arch.* **2023**, *48*, 1005–1012. [CrossRef]