EO Big Data Connectors and Analytics for Understanding the Effects of Climate Change on Migratory Trends of Marine Wildlife

Z. A. Sabeur, G. Correndo, G. Veres, B. Arbab-Zavar, J. Lorenzo, T. Habib, A. Haugommard, F. Martin, J.-M. Zigna, G. Weller

EO Big Data Connectors and Analytics for Understanding the Effects of Climate Change on Migratory Trends of Marine Wildlife

Z. A. Sabeur[(1)], G. Correndo[(1)], G. Veres[(1)], B. Arbab-Zavar[(1)], J. Lorenzo[(2)], T. Habib[(3)], A. Haugommard[(3)] and F. Martin[(3)], J-M. Zigna[(4)] and G. Weller[(4)]

{zas,gc,gvv, baz}@it-innovation.soton.ac.uk

(1) University of Southampton IT Innovation Centre, Department of Electronics and Computer Science, Faculty of Physical Sciences and Engineering, Southampton, United Kingdom

(2) Atos, Spain

(3) Atos, France

(4) Collecte Localisation Satellite (CLS), France

Abstract. This paper describes the current ongoing research activities concerning the intelligent management and processing of Earth Observation (EO) big data together with the implementation of data connectors, advanced data analytics and Knowledge Base services to a Big Data platform in the EO4Wildlife project (www.eo4wildlife.eu). These components support on the discovery of marine wildlife migratory behaviours, some of which may be a direct consequence of the changing Met-Ocean resources and the globe climatic changes. In EO4wildlife, we specifically focus on the implementation of web-enabled advanced analytics web services which comply with OGC standards and make them accessible to a wide research community for investigating on trends of animal behaviour around specific marine regions of interest. Big data connectors and a catalogue service are being installed to enable access to COPERNICUS sentinels and ARGOS satellite big data together with other in situ heterogeneous sources. Furthermore, data mining services are being developed for knowledge extraction on species habitats and temporal behaviour trends. Also, high level fusion and reasoning services which process big data observations are deployed to forecast marine wildlife behaviour with estimated uncertainties. These will be tested and demonstrated under targeted thematic scenarios in EO4wildlife using a Big Data platform a cloud resources.

1 Introduction

EO4wildlife brings large number of multidisciplinary scientists such as marine biologists, ecologists and ornithologists around the world to collaborate closely together while using European Sentinel Copernicus Earth Observations more efficiently.

In order to reach such important capability, an open service platform and interoperable toolbox is being designed and implemented. It offers data processing services that can be accessed by scientists to perform their respective research. The platform front end will be easy to use, access and it offers dedicated services that will enable scientists' process their geospatial environmental stimulations using Sentinel Earth Observation data and other observation sources. Specifically, the EO4wildlife platform will enable the integration of Sentinel data, ARGOS archive databases and real time thematic databank portals, including Wildlifetracking.org, Seabirdtracking.org, and other Earth Observation and MetOcean databases; locally or remotely, but simultaneously. EO4wildlife research specialises in the intelligent big data processing, advanced analytics and a Knowledge Base for wildlife migratory behaviour and trends forecasting. The research is leading to the development of web-enabled open services using OGC standards for sensor *Observation and Measurements* and data processing of heterogeneous geospatial observation data with estimated uncertainties. EO4wildlife designs, implements and validates various scenarios based on real operational use case requirements in the field of marine wildlife migrations, habitats and behaviour.

2 Global Architecture Overview

The EO4wildlife system is hosted in a SparkInData platform, which offers a set of core services for data discovery, data ingestion, process integration and execution. The SparkInData Platform, also known as *Smart Elastic Enriched Earth Data* (SEEED), is a generic platform which provides an EO data dedicated Cloud platform, infrastructure and services. Furthermore, the platform is organized under three functional zones, as shown in Figure 1 below. These include: *1- Storage zone for mutualized storage capabilities; 2- Compute zone for mutualized intensive computing; and 3- Service zone for processing services*. Furthermore, the platform infrastructure services are provided by the Big Data Helix Nebula platform. Slipstream is used at a *Platform as a Service* (PaaS) level. The PaaS is provided under a cloud computing services environment which enables developers run, test and manage their own applications while processing Big EO Data and analytics for extracting marine species migratory routes with respect to Ocean fronts geospatial and temporal trends. Specifically, PaaS is based on Google's Kubernetes (K8S) open software with an augmented dedicated SparkInData *Service Management Layer*. The latter is responsible for deploying applications on SaaS mode and managing them with auto-scaling, load balancing, monitoring or decommissioning upon request by application owners.

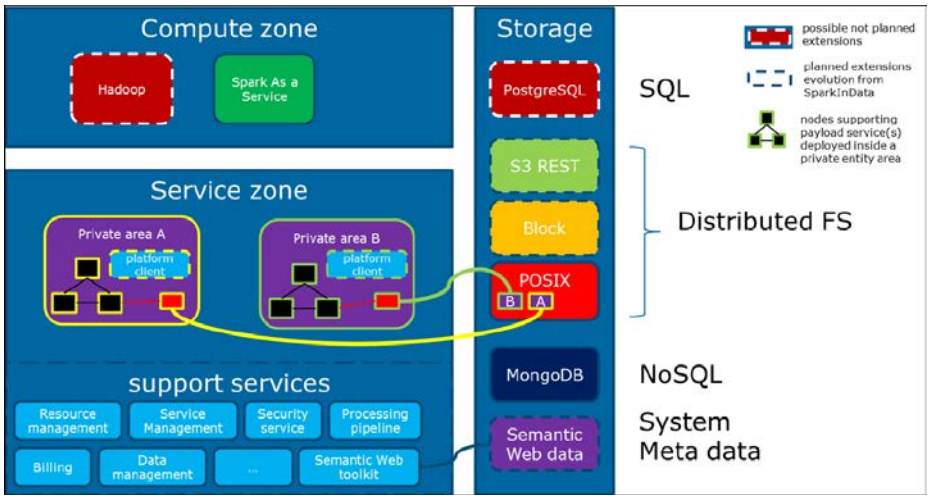


Fig. 1. Global Big Data Platform Architecture Overview

In addition to the above, the SparkInData platform components consist of the following:

***Security Service**:* It registers users, their roles and rights to ensure their authentication and access control to the platform
Processing Pipeline: It controls and monitors the chaining of the data analytics Web Processing Services (WPS)
***Data management service**:* It ensures the "import" into the system of various data sets which are generated by data providers or new analytics into the system
Service management: This enables the creation and control of applications deployment and their scalability of operations on the platform
Resource Management: It controls resources deployment, their scalability and operations on the platform
Data Storage service: This service ensures and secures persistent data storage into the system.
Semantic Web Toolkit: It ensures linked data storage, access to RDF resources and mechanisms which define access control policies for graph stores
Market Place: It provides a common place for information exchange for publishing application services outputs.
Billing service: It controls how application services can be purchased from the *Market Place* where payment can be performed
***Spark as a Service**:* It executes and controls Spark Jobs through a dedicated web service

2.1 Knowledge Base and Big Data Analytics

The Knowledge Base services are being developed in support of the overall EO4wildlife platform architecture. They are integrated with it for the support of the deployment of the big data analytics services to the Platform. The overall aim of the Knowledge Base services is to enhance the meta-data support provided by OGC

standards in order to employ data semantics and interoperability at the data access service level. The ontologies developed within this module aim at covering the gap between data producers (e.g. Argos, Copernicus or the animal tracks data owners) and the data consumers (e.g. the scientists which will develop the workflows).

This is achieved through the provision of a common overarching representation of the heterogeneous entities that access the Big EO data sources and analytics. The Knowledge Base architecture is made of various services functionalities which are exposed to the platform as REST services. The services are deployed as a separate Docker container and it is supported by the Virtuoso triple store instance which is included in the SEEED platform.

2.2 Big Data Analytics

In order to provide proofs of concept of the EO4wildlife platform, a number of work-flows based on specific domain studies of marine animal tracks are being developed. The more mature studies are presented in the following Sections.

2.2.1 Atlantic Bluefin Tuna Application Scenario

Atlantic Bluefin tuna (ABFT henceforth) is a highly migratory species which tolerates wide ranges of environmental conditions [1] in the Atlantic Ocean and Mediterranean Sea. In this section, we present initial attempts to correlate ABFT tracking data and environmental variables to identify different pattern of ABFT behaviour using Ecological Niche Modelling. The methodological steps undertaken to identify ABFT habitat preferences during different types of behaviour are inspired by [2] and depicted in **Fig. 3.**

Fig. 2. Workflow of Atlantic Bluefin tuna application

For **animal tracks pre-processing** the following services are offered: animal tracks reconstruction, discarding the location on the land and redundancy filtering which removes duplications in the tracks and any relocation points on the same day separated by less than 2.3 km. **EO (Earth Observation) data pre-processing** includes bringing all environmental variables to the same spatial and temporal resolutions, smoothing filters to recover spatial missing data, calculating 3-7 days composites for variables with temporally missing data and performing additional operations on data such as gradients calculations.

Specifying habitat of ABFT. Two patterns of behaviour can be observed for ABFT: spawning and feeding [1]. In the Mediterranean Sea, according to the literature spawning takes place between middle of May and middle of July months, while feeding is between mid-July and mid-September months, in the Mediterranean Sea or the North

Atlantic Ocean. The following environmental variables were identified as relevant to influencing habitat utilization [2]: Daily sea surface temperatures (SST) and chlorophyll concentration (CHL) which can be used to calculate respective gradients and fronts; bathymetry; CO_2 net Primary Production (PP); daily sea surface height anomaly (SSHa), ocean currents (eddies) and wind speed at sea surface level. Additionally, several previous research papers [2] reported the specific environmental conditions which are favoured by ABFT for spawning and feeding activities. For spawning habitat, Bluefin tuna prefers warm waters of the Mediterranean Sea (SST in the range 20 to 25.5ºC) with increasing SST over several weeks, relatively low levels of CHL, intermediate levels of Eddy Kinetic Energy (EKE) and preferable range for SSHa. While for feeding habitat, ABFT prefer to locate in the vicinity of chlorophyll frontal features and higher levels of concentrations, wide range of SST and immediate levels of PP. These observations and analyses show that it should be possible to distinguish between feeding and spawning behaviors of ABTF when environmental variables are added to modelling.

Environmental analysis. Our environmental analyses started with investigating visual correlation between ABFT tracks and environmental variables. It could be visually observed that some mean values of environmental variables have different ranges for spawning and feeding such mean SST (**Fig. 4**) and CHL (**Fig. 5**). Additionally, environmental variables with different ranges for spawning and feeding are SSHa, CO_2 Net PP and Eddy Kinetic Energy (not shown due to space limitation). Note that though tracks for feeding (Adriatic Sea) and spawning (the Mediterranean Sea) do not cover exactly the same area, the difference in ranges between different behaviors are consistent with other findings in the Mediterranean Sea, i.e. This is due to different behavior of ABFT (spawning and feeding) to a large extent rather than due to different areas. However, further investigation will include more tracks from the Mediterranean Sea for both spawning and feeding to confirm these early findings.

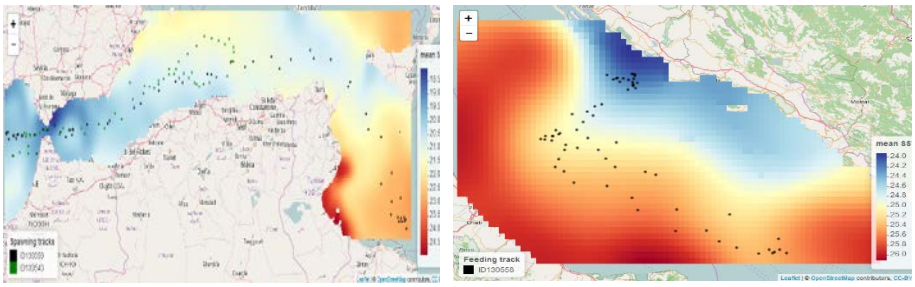


Fig. 3. Mean SST ranges for spawning (right) and feeding (left) tracks

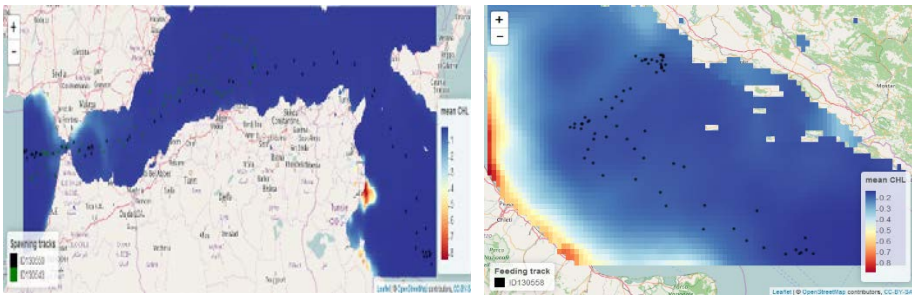


Fig. 4. Mean CHL for spawning (right) and feeding (left) tracks.

Further, the link between environmental variables and ABFT presence data will be analysed using a cluster analysis [5] and/or histogram approach. The cluster analysis is suitable for separating different behaviors that occur in distinct environments, for example feeding and spawning, or suitable and unsuitable habitats. K-means clustering can be explored for identifying the relevant thresholds for ecological variables (the most suitable for spawning), and overall environmental envelopes for feeding. When cluster analysis is used to define the relevant thresholds, the 15^{th} and 85^{th} percentile values can be used since they represent relatively extreme environmental boundaries while rejecting the potentially misclassified distribution tails [2]. Alternatively, a histogram approach may identify thresholds for environment variables with adding some uncertainty in the calculations due to the nature of tracking data. Since tracking devices report only where particular set of fish went, while we do not know the areas with no fish or indeed areas where other untracked fish went. We can overcome this difficulties to some extent by considering a presence/availability paradigm rather than a presence/absence model and generating potential tracks using a Null model movement paths based approach.

Habitat Modelling. Once the threshold values for environmental variables are set, the specific ecological niche of ABFT will be defined for feeding and spawning. An ecological niche model using environmental envelopes can be used to predict the daily suitability of cells within habitat for ABFT feeding and spawning on a given [0, 1] scale. The favorable habitat for each behavior are cells that meet all the suitable ranges of selected variables. This will be the next step of our services development.

Performance evaluation. The performance evaluation can be challenging due to a small number of tracks available and observations covering only presence of fish tracked. However we can compare our potential results with findings reported in the literature and by computing the distance between the presence data and the closest favourable habitat (3-day composite) for available tracks

2.2.2 Marine Turtle Application Scenario

Another exemplar data processing workflow which is being evaluated under the EO4wildlife platform as a proof of concept is the marine turtles application scenario. It is based on the workflow described by Pikesley et al.[3]. 21 female Olive ridley turtles

are tracked between 2007 and 2010 in the south-east of the Atlantic ocean near the west-African coast [4]. The aim is to describe the observed and potential post-nesting habitat for these species in the region. This will also be important for the fishing industries in the region to become aware of areas with the presence of turtles that could potentially lead to their unnecessary bycatch. A similar approach is taken in [3], where 32 adult loggerhead turtles are tracked in the eastern part of the Atlantic. This work investigated how the predicted habitat may alter following climate change. **Fig. 6** depicts the workflow which is implemented within EO4wildlife's sea turtle scenario. It is based on the analysis described by Pikesley et al. in [3].

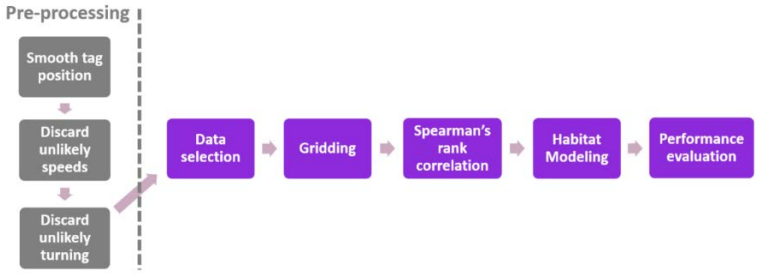
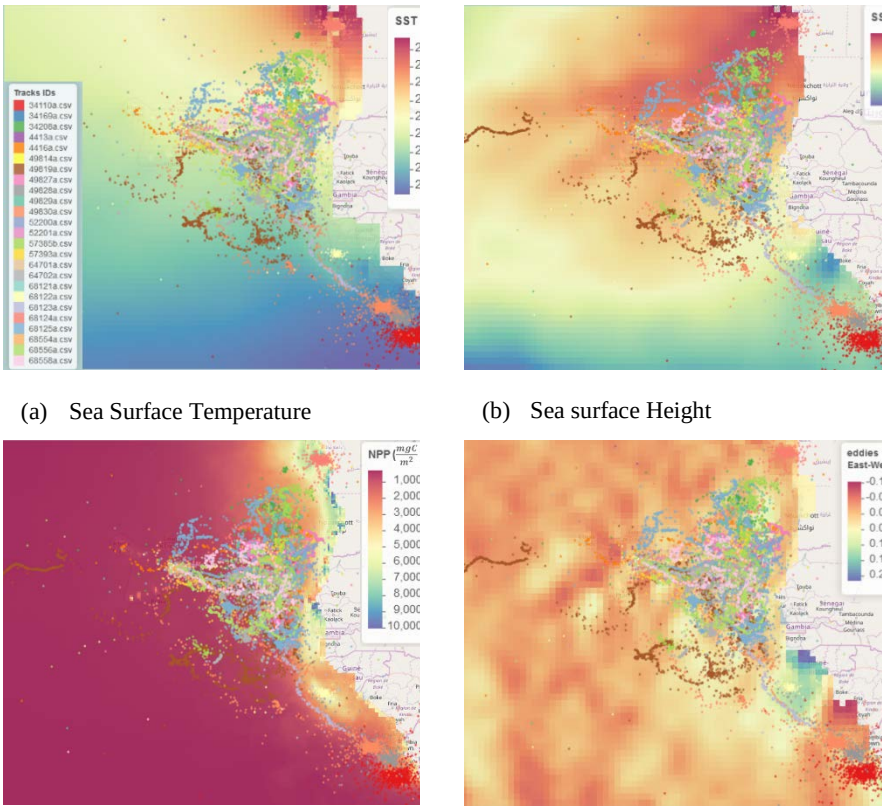


Fig. 5. Marine Turtle workflow

Other approaches describing the marine turtles' behaviour have also been considered and may be looked upon for further workflows in the marine turtle scenario. These workflows look at different sea turtle behaviors including: nesting activity and clutch frequency [5] and post-nesting migration and foraging behavior [6]. Habitat modelling via ecological niche modelling is one of the more systematic ways of analysing species distribution under climatic changes. Other climate change scenarios should be considered to make the causal projections of environmental variables. A Spearman's rank correlation test will then be calculated for each paired environmental variables. These environmental variables are sampled from long-term mean values at the location of track points. Twenty five tracks of adult loggerhead sea turtles are being evaluated during their post-nesting movements near the west coast of Africa. This data includes samples from Aug 2004 to Dec 2009, where in some of the turtles are tracked for a short time and some for longer periods. The considered environmental variables include: *1) Sea Surface Temperature; 2) Bathymetry; 3) Sea surface Height (Absolute Dynamic Topography); 4) Net Primary Production; 5) Current Velocity; and 6) Eddies.* These environmental conditions were used and based on their previous reported correlations with marine turtle relocations [3], [4], [6]. **Fig. 8** shows the long-term mean of some of these variables that are superimposed with the species track points.



(a) Sea Surface Temperature (b) Sea surface Height

(c) Net Primary Production (d) Eddies East-West

Fig. 6. Superimposed tracks points with environmental variables

2.3 Web-Enabled Data Analytics Services

The data analytics based approaches of Section 2.2 are being wrapped as open web services. They have been specifically designed with fundamental capabilities that are exposed as OGC compliant WPS services [7]. This will also ease their integration in the overall SEEED big data platform architecture and increase their reusability between applications. These services have been implemented as Docker containers [8] whose instantiation is managed via the Kubernetes API. The adoption of Docker technologies ensures that the services' implementations are self-contained and their dependencies explicitly declared by means of Docker files. This is for the purpose of their invocation accordingly with respective to given input parameters to the data analytics processing algorithms. The deployment structure of such services in depicted in **Fig. 9** below:

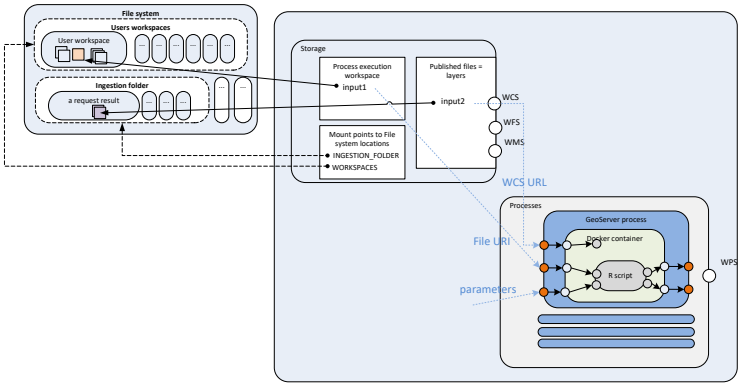


Fig. 7. Data analytics services deployment for the EO4wildlife big data platform architecture

The data analytics based algorithms which are deployed in the platform primarily include mining, machine learning, data fusion and reasoning methods on marine species behaviour which correlates with geospatial and temporal trends of MetOcean fronts. The resulting open web processing services are grouped in three main categories: data pre-processing and aggregation, data mining services and high level fusion services. The high level fusion services are those which group all the intelligent post-processing of data analytics, forecasting and reasoning through knowledge modelling.

2.4 Data Pre-processing and Aggregation

This category contains the services for the pre-processing, cleaning and aggregation of the data prior to data analytics. The pre-processing of geospatial data sets is an important step when dealing with potentially incorrect information such as non-plausible animal positions. These services allow the recognition and removal of all data elements which are clearly unrealistic given the knowledge of the domain (e.g. the animal is not capable of travelling at such velocities, or producing abrupt changes of directions or; again travelling inland). Moreover the pre-processing services allow the filling of missing data values and accommodation of different data grids using interpolations which are not directly collected and represented in the data sets. The aggregation service allows the reconciliation of data which represented with different spatial or temporal resolutions. It also provides functionalities to sample environmental observations and aggregate them into the required grid resolution for input to the niche modelling algorithms. Under this category, we also included services which process animal tracks to provide grouping of tracks in trips or gridding a number of tracks to compute and analyse species population distributions.

2.5 Data Mining Services

These services process animal tracks and satellite marine observations in order to model animals' use of space and correlate it with environmental observations. This is further subdivided in two sub-categories of services: Animal track based services and

statistical environmental services. These respectively analyse animal tracks for estimating the species home range and foraging grounds; and the statistical relevance of environmental observations in modelling animals' presence.

2.6 High Level Fusion Services

This category contains services that make use of multiple data sources to better estimate animals' position, behaviour and modelling animals' habitats. This category includes the Track & Lock service which enables the estimation of submarine trajectories for animals equipped with pop-up or archival tags, Change-point Analysis and various Habitat modelling techniques.

3 Acknowledgements

This research is partly funded by the European Commission under H2020 Grant Agreement number 687275. We are also grateful to the European Commission for giving us access to Copernicus Satellite observation data in this project. Access to Argos satellite databases through our consortium partnership is also acknowledged.

References

[1] H. Arrizabalaga *et al.*, 'Global habitat preferences of commercially valuable tuna', *Deep Sea Research Part II: Topical Studies in Oceanography*, vol. 113, pp. 102–112, 2015.

[2] J.-N. Druon *et al.*, 'Habitat suitability of the Atlantic bluefin tuna by size class: An ecological niche approach', *Progress in Oceanography*, vol. 142, pp. 30–46, 2016.

[3] S. K. Pikesley *et al.*, 'Modelling the niche for a marine vertebrate: a case study incorporating behavioural plasticity, proximate threats and climate change', *Ecography*, vol. 38, no. 8, pp. 803–812, 2015.

[4] S. K. Pikesley *et al.*, 'On the front line: integrated habitat mapping for olive ridley sea turtles in the southeast Atlantic', *Diversity and Distributions*, vol. 19, no. 12, pp. 1518–1530, 2013.

[5] Al. F. Rees, A. Al-Kiyumi, A. C. Broderick, N. Papathanasopoulou, and B. J. Godley, 'Conservation related insights into the behaviour of the olive ridley sea turtle Lepidochelys olivacea nesting in Oman', *Marine Ecology Progress Series*, vol. 450, pp. 195–205, 2012.

[6] P. Chambault *et al.*, 'The influence of oceanographic features on the foraging behavior of the olive ridley sea turtle Lepidochelys olivacea along the Guiana coast', *Progress in Oceanography*, vol. 142, pp. 58–71, 2016.

[7] M. Mueller and B. Pross, 'OGC WPS 2.0 Interface Standard', *OpenGeospatial Consortium Inc, OGC*, pp. 14–65, 2015.

[8] D. Merkel, 'Docker: Lightweight Linux Containers for Consistent Development and Deployment', *Linux J.*, vol. 2014, no. 239, Mar. 2014.