Deliverable 1.4 Final data management plan

Deliverable Contributors:	Name	Organisation	Role / Title
Deliverable Leader	Jozef Skakala	PML	Project coordinator
Contributing Author(s)	Pierre Brasseur	UGA/CNRS	PI
	Laurent Bertino	NERSC	PI
	Gianpiero Cossarini	OGS	PI
	Lars Nerger	AWI	PI
Reviewer(s)	Jessica Heard	PML	Project manager
	Stefano Ciavatta	Mercator Ocean	Advisory Board
Final review and approval	Jozef Skakala	PML	Project coordinator

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101004032.

Document History:

Release	Date	Reason for Change	Status	Distribution
1.0	20/12/23	Revised	Released	Public
0.0	20/11/23	Initial document	Released	Internal

To cite this document

This document is a public deliverable. It should be cited as Skakala J, Brasseur P, Bertino L, Cossarini G and L Nerger, 2023. Deliverable D1.4: Final data management plan, EU H2020 SEAMLESS project, grant agreement No 101004032. Doi; 10.5281/zenodo.10405912

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101004032.

TABLE OF CONTENTS

1. Executive Summary 4
2. Scope 4
3. Data summary 4
4. FAIR data 7
   4.1 Making data findable, including provisions for metadata 7
   4.2 Making data openly accessible and interoperable 7
5. Other research output 8
   5.1 The SEAMLESS prototype: EAT 8
   5.2 The code for stochastic parameterizations to generate ensembles in 1D or 3D MFC configurations 9
   5.3 The code for ensemble evaluation 9
   5.4 The code of the ensemble data assimilation systems 9
6. Allocation of resources 10
7. Data security 10
8. Ethical aspects 11
9. Appendix: Data Register 12

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101004032.

1. Executive Summary

SEAMLESS participates in the Open Research Data Pilot (ORDP) and follows the FAIR (Findable, Accessible, Interoperable, Re-usable) data paradigm and this is reflected in this Final Data Management Plan (DMP).

This DMP details the management of data generated within the SEAMLESS project. It covers:

- what has been generated,
- how data have been handled during and will be after the project,
- who owns a data set and with whom it is shared,
- how sharing of data within and outside the project is organised,
- what formats, metadata and standards the data have adhered to.

The SEAMLESS data sets consist of numerical simulation outputs generated by a range of marine ecosystem models and data assimilation systems. The code of the models and data assimilation systems themselves were a major outcome disseminated by the project.

This DMP describes relevant procedures put in place during the project to adhere to principles of accessibility and interoperability, ensuring that both the data and codes developed by the project are useful to project partners and to other stakeholders.

2. Scope

This document, comprising the Data Management Plan, is intended for internal and external use, describing the mechanisms that SEAMLESS has put in place to ensure all public data follow the FAIR (Findable, Accessible, Interoperable, Re-usable) data management principles.

The Data Management Plan has been updated as a “live” document during the lifetime of the project. An Initial Data Management Plan was delivered at month six of the project (SEAMLESS Deliverable 1.3) The current document presents the status and planning at month 36 of the three-year project.

3. Data summary

The overall objective of SEAMLESS is to provide the Copernicus Marine Service with new capabilities to deliver indicators of climate-change impacts and food security in marine ecosystems.

To achieve this objective, SEAMLESS has produced both model code and data of marine physical and biogeochemical variables from model simulations.
These simulations cover the European regional seas matching the domains of the Monitoring and Forecasting Centres (MFCs) of the Copernicus Marine Service, as well as the North Atlantic Ocean (see Figure 1). In addition, one-dimensional simulations (water column profiles) have been delivered in correspondence of data-rich monitoring sites for comparison with field observations (Figure 1).

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 101004032.

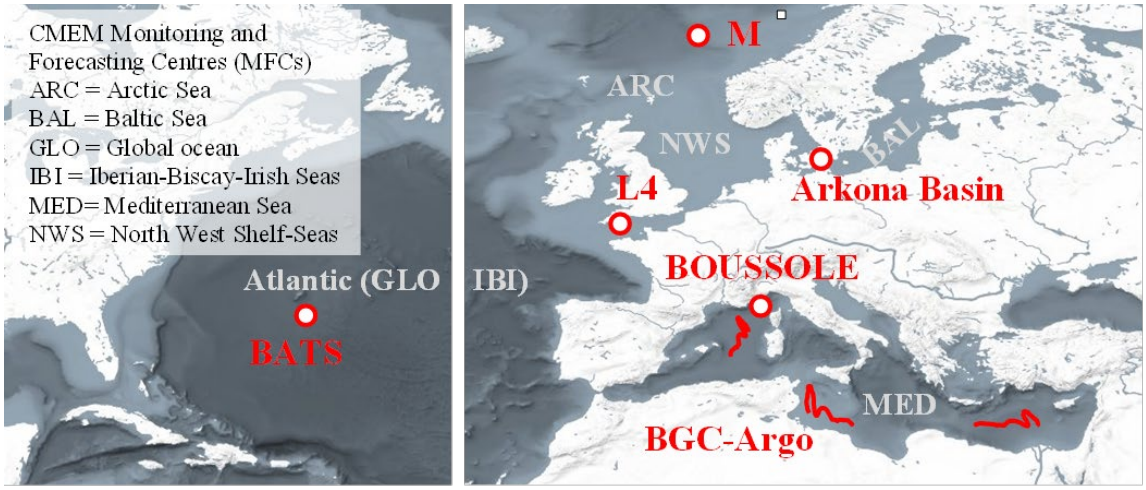

Figure 1. SEAMLESS has produced model output data for the six three-dimensional domains of the Copernicus Marine Service Monitoring and Forecasting Centres (MFCs) listed and represented in the map. In addition, the project will deliver one-dimensional simulations (water column profile) in correspondence of data-rich monitoring sites highlighted by the red circles. The red lines represent the trajectories of the biogeochemical-Argo floats (BGC-Argo) simulated with Lagrangian configurations of the biogeochemical model of the Mediterranean Sea.

The project has produced model outputs with the primary aim of assessing the performance of the data assimilation methods and biogeochemical models developed or advanced in the project, through comparison with observations available in the literature or provided by project stakeholders.

We expect that the improved simulations of the biogeochemical variables and indicators can be exploited by project stakeholders to:

1) understand better the functioning of marine ecosystems (e.g. carbon export, phytoplankton community succession, productivity);
2) assess the health of the ecosystem in relation to, e.g., acidification, deoxygenation and eutrophication indicators;
3) evaluate alternative monitoring strategies (e.g. deployment of biogeochemical-Argo floats and/or gliders);
4) consider upgrades of the current modelling and assimilation systems of the Copernicus Marine Service MFCs (e.g. using hybrid assimilation methods rather than variational only in the Mediterranean and North-West European Seas).

The outputs include primarily the ecosystem indicators that are targeted by SEAMLESS in all the simulated domain, or the state variables needed for their easy computation (see Table 1). Additional biogeochemical variables have been produced by the partners in relation to the observations available for model assessment in their target region (e.g. nutrients). Physical data has also been produced as by-products of the coupled biogeochemical-physical model simulations.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101004032.

Table 1 List of biogeochemical **indicators** which has been delivered as primary output of the SEAMLESS model simulations

1) Particulate organic carbon (POC)
2) Trophic efficiency (as zooplankton normalized by phytoplankton concentrations)
3) Primary production
4) pH
5) Oxygen concentration
6) Phytoplankton functional types (PFT; carbon biomass)
7) Phytoplankton phenology (initiation, peak and duration of the blooms, as quantified by the chlorophyll concentrations)

The biogeochemical datasets produced by SEAMLESS are listed and described in Tables 2 relative to the Copernicus Marine Service MFCs regions, and in Table 3 relative to the data-rich sites.

Table 2 List of the model output datasets that were produced by SEAMLESS for the Copernicus Marine Service MFC regions shown in Figure 1. The primary content of the datasets will be the biogeochemical indicators listed in Table 1.
The datasets were be produced as three-dimensional fields of daily or monthly values. All the datasets are in NetCDF format.

Partner	Region	Model	Size of model output (Tb)
PML	NWS	NEMO-ERSEM	1.5
AWI	BAL	NEMO-ERGOM	0.6 TB per simulation year
IGE	GLO/ATL	NEMO-PISCES	37.0
NERSC	ARC	HYCOM-ECOSMO	0.05
OGS	MED	NEMO-BFM	0.5Tb per year simulated

Table 3 List of the model output datasets that were produced by SEAMLESS for the data rich sites shown in Figure 1. The primary content of the datasets will be the biogeochemical indicators listed in Table 1.
The datasets were be produced as one-dimensional profiles of daily or monthly values. All the datasets are in NetCDF format.

Partner	Region	Model	Size (Gb)
PML	NWS	GOTM-ERSEM	10
AWI	BAL	GOTM -ERGOM	0.1
IGE	GLO/IBI	GOTM -PISCES	30
NERSC	ARC	GOTM -ECOSMO	0.1
OGS	MED	GOTM -BFM	10

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101004032.

4. FAIR data

4.1 Making data findable, including provisions for metadata

All SEAMLESS research data were curated according to the 'FAIR' principle, i.e. to be Findable, Accessible, Interoperable and Re-usable. In the following, a short overview is given of the building blocks and guiding principles to reach this goal.

Any documentation such as model description and simulation set-up for final datasets were linked within the data register (see Section 4.2) and a copy has been kept in the SEAMLESS project management portal. Documents for public dissemination were made publicly available in Zenodo and gitHUB repositories.

By default, all data generated in SEAMLESS are openly available (see Data Access, below), with the exception of unprocessed, uncorroborated data produced in test simulations if these have no value to the user. Data contributed from external sources are the exception to this rule. In such cases, data ownership and licensing will govern whether dissemination beyond SEAMLESS is possible. Reference to existing FAIR data sources is to be preferred over duplication.

The metadata used to describe datasets made available followed the standards of the Open Geospatial Consortium (OGC). As a common ontology the CF conventions (cfconventions.org) have been followed. These metadata conventions ensure that data are identifiable using appropriate search terms and key words.

4.2 Making data openly accessible and interoperable

Data generated as part of SEAMLESS are free of cost to the user. Data access restrictions and intellectual property rights, however, remain as set by the dataset creator/owner where applicable. Unless specified, all data were treated as FAIR open data. In practise, the following data access levels are guaranteed:

- open access, not requiring registration, providing access to data identified as open without license restrictions;
- limited access, requiring registration, providing access to open data as well as data sets with a limited license for use (e.g. non-commercial, accrediting ownership, delayed release etc);
- restricted access, requiring registration, providing access to data owned by the user and any data sets to which this specific user has been granted access.

The model output datasets produced within the project have been stored in the servers of the partners producing the data. The data are made freely available to users, on request.

Model output data will be kept available for a minimum of three years after the end of SEAMLESS. Beyond this period, e.g. if the service should no longer be deemed useful or sustainable, data will be archived at a secure open access location, insofar as data licensing permits.

Requests to remove a data set from the SEAMLESS services can be submitted to the Coordinator and will be handled in a manner equivalent to the GDPR for personal data.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101004032.

Datasets that meet the criteria for dissemination are listed in the appendix of the DMP.

5. Other research output

SEAMLESS research outputs also included software and model code, besides the model products described in the above Sections 1 and 2. Most notably, SEAMLESS delivered:

1. The SEAMLESS prototype: EAT;
2. The code for stochastic parameterizations to generate ensembles in 1D or 3D MFC configurations;
3. The code of the new ensemble data assimilation systems.

Which are briefly described in the following subsections (please see the project deliverables for the thorough descriptions of the code and their availability)

5.1 The SEAMLESS prototype: EAT

The project delivered the SEAMLESS prototype: the Ensemble and Assimilation Tool (EAT), which is a flexible and extensible software package that enables data assimilation of physical and biogeochemical variables in a one-dimensional water column. EAT itself is the Deliverable D2.4 (nature "Other") of SEAMLESS and was supported by a documentation report. The EAT source code is available from https://github.com/BoldingBruggeman/eat. It includes compatible versions of GOTM and FABM as submodules. FABM in turn includes numerous biogeochemical models, including ECOSMO. In addition, an adapted version of PDAF is included in the EAT source code. The 3D variational algorithms that have been implemented in PDAF during the SEAMLESS project are also publicly available in the PDAF releases starting from Version 2.0. They are available via the PDAF website https://paf.awi.de and at https://pdaf.awi.de/PDAF/PDAF.

The example applications use three biogeochemical models that are hosted externally:

- PISCES: https://github.com/BoldingBruggeman/fabm-pisces
- BFM: https://github.com/inogs/bfmforfabm
- ERSEM: https://github.com/pmlmodelling/ersem

Documentation is available at https://github.com/BoldingBruggeman/eat/wiki

Binary installation packages are available via the package manager system conda and can be installed with a single line on Linux, Mac and Windows systems that have the Anaconda Python distribution:

```
conda create -n eat -c bolding-bruggeman -c conda-forge eatpy
```

If necessary, Anaconda can itself be installed with a single command, without requiring administrator/root permissions).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101004032.

5.2 The code for stochastic parameterizations to generate ensembles in 1D or 3D MFC configurations

SEAMLESS delivered codes to generate ensembles with NEMO and HYCOM, which are the building blocks of all CMEMS modelling systems. This code was developed as part of Task 3.1 of WP3 and was described in documentation attached to Deliverable 3.3 (D3.3, Type: Code). The D3.3 codes were delivered to the SEAMLESS stakeholders in open-access repositories

In particular, the stochastic coupled NEMO-PISCES version has been developed in the global setup implemented by the GLO MFC system (version r4.0-HEAD.r13720, provided by Mercator Ocean International) as part of Task 3.1. The resulting probabilistic code is available here: https://zenodo.org/record/6303007.

The probabilistic formulation of HYCOM-ECOSMO relies on a method for generating two-dimensional pseudo random fields with a specific mean, variance, and co-variance as described in Evensen (1994). The code can be accessed here: https://zenodo.org/record/5606973#.YeVsMmDTVqs. The routines for creating perturbations in 3D (x,y,z) or 2D (x,y,t) are described here: https://github.com/lbertino/Random_2D_perturbations.

5.3 The code for ensemble evaluation

SEAMLESS has developed a toolbox to assess the quality and relevance of model-generated ensembles and to provide probabilistic evaluation scores of ensemble assimilation methods (see the project Deliverable D3.1: Brasseur P, Brankart J.M., Popov M, (2022). D3.1 Code for ensemble evaluation in prototype. Deliverable report of project H2020 SEAMLESS (grant 101004032.), doi: 10.5281/zenodo.6390306). The package has been conceived in order to be easily plugged into existing software (including the SEAMLESS prototype described in section 4.1), or applicable to existing ensemble-based CMEMS assimilation systems ,as well as other external prediction systems (e.g. the OceanPredict community). The user's guide (see the Appendix of D3.1) was written with the objective of presenting the tools in a consistent and "demystified" way in order to facilitate their learning, while avoiding the more complex mathematical formulations and technical details that are otherwise available in the referenced publications.

The code is publicly available here: https://github.com/brankart/ensdam

5.4 The code of the ensemble data assimilation systems

SEAMLESS developed new or updated existing ensemble data assimilation systems for all the Copernicus Marine Service domains, as a major outcome of the project. The assimilation systems are described extensively in deliverable D3.4 (Brasseur P. *et al.* (2022). D3.4 Observability of the target indicators in the 3D CMEMS MFC systems. Deliverable report of project H2020 SEAMLESS (grant 101004032.). doi: 10.5281/zenodo.7584865) and their applications in deliverables D4.1, D4.2, D5.1 and D6.1 of SEAMLESS.

The code of the assimilation systems has either been integrated in the prototype described in Section 4.1, or is available in public git repositories or in repositories hosted by the project partners: see Table 4.1.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101004032.

System	DA method	Code repository	Availability	Contact
NWS SEAMLESS	Hybrid Ensemble/NEMOVar	MFC Met Office MonSOON HPC	available upon request	Jozef Skakala jos@pml.ac.uk
IBI/GLO SEAMLESS	Stochastic Ensemble Filter	https://github.com/brankart/ensemble-ocean-colour https://github.com/brankart/sesam https://github.com/brankart/ensdam	public	Jean-michel.brankart@univ-grenoble-alpes.fr
MED SEAMLESS	SEIK	https://github.com/inogs/ogstm	public	Simone Spada sspada@ogs.it
BAL SEAMLESS	LESKTF & Hybrid Filter LKNETF	https://github.com/PDAF/NEMO-PDAF		Lars Nerger lars.nerger@awi.de
ARC SEAMLESS	DEnKF	https://github.com/nansencenter/enkf-topaz	public	Tsuyoshi Wakamatsu tw@nersc.no

6. Allocation of resources

The development and maintenance of the SEAMLESS project management portal was the responsibility of PML, who will continue to maintain access for at least three years beyond the end of the project.

Management of the repositories of the datasets in the partners' servers is responsibility of the partners.

7. Data security

To safeguard the final model output datasets, backups are made at the site where they are hosted.

Copies of test simulation data will, in general, not be backed-up by SEAMLESS after the final datasets are delivered.

A number of data repositories have been set up to safeguard specific project outputs, such as software (such as the PML GitHub repository), publications (such as the Plymouth Marine Science Electronic Archive, PlymSEA), sensitive data (internal project management databases, which are backed up onto a separate PML server plus an external additional backup copy).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101004032.

8. Ethical aspects

Ethical aspects are mainly relevant for data of a personal nature. These data were treated according to the ethics procedures laid out in the Ethics section of the GA, in summary these procedures cover the following aspects:

- Details on the informed consent procedures for the participation of humans;
- Templates of the informed consent forms;
- Templates of the user and stakeholder questionnaires.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101004032.

9. Appendix: Data Register

Table 9.1. Template with description of the data register

Organisation	**Dataset reference & Name**	**Dataset description/outline**			**Standards & metadata**	**How will data be shared**	**Software/protocol required for sharing**	**Data access policy (open/locked/ partial - give details, e.g. embargo time)**	**Location of storage**	**Documentation (deliverable)**
Name of organisation providing the data. Also reference any other ownership, i.e. if you have bought commercial data and you have rights to use but must attribute	*A unique reference label*	*Simple description of the dataset, try to include as much information as possible*	*Spatial Resolution & extent*	*Temporal resolution and extent*	*Any standardised metadata that accompanies the dataset*	*List data services or custom websites*	*List the protocols available for data access*	*Data policy, such as groups that can use, whether it is only accessible to project partners or whether there is a time based embargo*	*Where and how data will be stored*	*The SEAMLESS deliverable where the model and/or dataset are described*

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101004032.

Table 9.2 Data register for the dataset of the North West Shelf (NWS) domain of the Copernicus Marine Service produced in SEAMLESS by PML.

Organisation	**Dataset reference & Name**	**Dataset description/outline**			**Standards & metadata**	**How will data be shared**	**Software/protocol required for sharing**	**Data access policy (open/locked/ partial - give details, e.g. embargo time)**	**Location of storage**	**Documentation (deliverable)**
PML	REA_NWS_2018 (Reanalysis 2018 North-West Shelf seas domain)	NEMO-ERSEM output dataset for physical and biogeochemical variables	Copernicus Marine Service MFC NWS domain, 4-dimensional 7km horizontal resolution, variable vertical resolution	Daily means	NetCDF-CF	Transferred to user on request	Scp, sftp	open	Data are stored on MonSOON storage system MASS	D6.1

Table 9.3 Data register for the dataset of the Baltic Sea domain of the Copernicus Marine Service produced in SEAMLESS by AWI.

Organisation	**Dataset reference & Name**	**Dataset description/outline**			**Standards & metadata**	**How will data be shared**	**Software/protocol required for sharing**	**Data access policy (open/locked/ partial - give details, e.g. embargo time)**	**Location of storage**	**Documentation (deliverable)**
AWI	CHLw+SSTw (2015 reanalysis in the Baltic Sea)	NEMO-ERGOM output dataset. Ensemble mean of physical and biogeochemical variables	Baltic Sea 1.8km horizontal resolution, 56 layers.	Daily snapshots at midnight	NetCDF-CF	Transferred to user on request	scp, sftp		Stored at HLRN	D6.1

Table 9.4 Data register for the dataset of the Mediterranean Sea (MED) domain of the Copernicus Marine Service produced in SEAMLESS by OGS.

Organisation	**Dataset reference & Name**	**Dataset description/outline**			**Standards & metadata**	**How will data be shared**	**Software/protocol required for sharing**	**Data access policy (open/locked/ partial - give details, e.g. embargo time)**	**Location of storage**	**Documentation (deliverable)**
OGS	Med24_2019_Ens_OC_BGCfloat (2019 reanalysis in the Mediterranean Sea)	OGSTM-BFM output dataset. Ensemble mean of biogeochemical variables	Mediterranean Copernicus Marine Service MFC N domain, 4-dimensional 4.5km horizontal resolution, variable vertical resolution	Daily mean	NetCDF-CF	Transferred to user on request	CINECA sftp	open	Stored at CINECA	D6.1

Table 9.5 Data register for the dataset of the Global (GLO) and Iberian-Biscay-Irish Seas (IBI) domains of the Copernicus Marine Service produced in SEAMLESS by UGA.

Organisation	**Dataset reference & Name**	**Dataset description/outline**			**Standards & metadata**	**How will data be shared**	**Software/protocol required for sharing**	**Data access policy (open/locked/ partial - give details, e.g. embargo time)**	**Location of storage**	**Documentation (deliverable)**
UGA	ENS_GLOBAL_2019	Global ensemble simulation (40 members), including physical and bigeochemical variables	Global domain, 1/4° resolution, 75 z-levels	5-day means globally, daily means for the Atlantic, 2019	NetCDF-CF	Transferred to user on request	Opendap server	open	Stored at IDRIS computing center	D3.4/D6.1
UGA	REA_BATS_2019, REA_PAP_2019, REA_Senegal_2019	Ensemble reanalysis (40 members) for BATS, PAP and Senegal regions	Regional domains (BATS, PAP, Senegal), 1/4° resolution, 32 z-levels	Daily mean,2019	NetCDF-CF	Transferred to user on request	Opendap server	open	Stored at IDRIS computing center	D6.1

Table 9.3 Data register for the dataset of the Arctic (ARC) domain of the Copernicus Marine Service produced in SEAMLESS by NERSC.

Organisation	**Dataset reference & Name**	**Dataset description/outline**			**Standards & metadata**	**How will data be shared**	**Software/protocol required for sharing**	**Data access policy (open/locked/ partial - give details, e.g. embargo time)**	**Location of storage**	**Documentation (deliverable)**
NERSC	REA_ARC_noprm_2019	Ensemble reanalysis (80 members) for the ARC MFC domain including biogeochemical variables	Regional domain (Lat 43.5° to 90°Lon -180° to 179.75°), 1/4° resolution, 40 z levels	Daily mean, 2019	NetCDF-CF	Transferred to user on request	sftp	open	Stored at Sigma2 NIRD data storage	D6.1