MAchinE Learning for Scalable meTeoROlogy and climate

Data Management Plan

Daniel Thiemert

www.maelstrom-eurohpc.eu

D4.2 Data Management Plan

Author(s): Daniel Thiemert (ECMWF)

Dissemination Level: Public

Date: 21/09/2021
Version: 1.0
Contractual Delivery Date: 30/09/2021
Work Package/ Task: WP4/ T4.1
Document Owner: ECMWF
Contributors: All Partners
Status: Final

MAELSTROM

Machine Learning for Scalable Meteorology and Climate

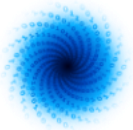
Research and Innovation Action (RIA)
H2020-JTI-EuroHPC-2019-1: Towards Extreme Scale Technologies and Applications

Project Coordinator: Dr Peter Dueben (ECMWF)
Project Start Date: 01/04/2021
Project Duration: 36 months

Published by the MAELSTROM Consortium

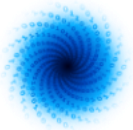
Contact:
ECMWF, Shinfield Park, Reading, RG2 9AX, United Kingdom
Peter.Dueben@ecmwf.int

The MAELSTROM project has received funding from the European High-Performance Computing Joint Undertaking (JU) under grant agreement No 955513. The JU receives support from the European Union's Horizon 2020 research and innovation programme and United Kingdom, Germany, Italy, Luxembourg, Switzerland, Norway



Contents

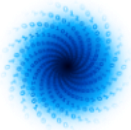
Tables



1 Executive Summary

The MAELSTROM Data Management Plan responds to the requirements of the H2020 Open Research Data Pilot to document which research data is being produced by the MAELSTROM project, in which format, and how it will be made available.

It has already identified data sets envisaged by the project but is only to be seen as an initial version which requires periodic updates to provide the necessary detail as it emerges.



2 Introduction

2.1 About MAELSTROM

To develop Europe's computer architecture of the future, MAELSTROM will co-design bespoke compute system designs for optimal application performance and energy efficiency, a software framework to optimise usability and training efficiency for machine learning at scale, and large-scale machine learning applications for the domain of weather and climate science.

The MAELSTROM compute system designs will benchmark the applications across a range of computing systems regarding energy consumption, time-to-solution, numerical precision and solution accuracy. Customised compute systems will be designed that are optimised for application needs to strengthen Europe's high-performance computing portfolio and to pull recent hardware developments, driven by general machine learning applications, toward needs of weather and climate applications.

The MAELSTROM software framework will enable scientists to apply and compare machine learning tools and libraries efficiently across a wide range of computer systems. A user interface will link application developers with compute system designers, and automated benchmarking and error detection of machine learning solutions will be performed during the development phase. Tools will be published as open source.

The MAELSTROM machine learning applications will cover all important components of the workflow of weather and climate predictions including the processing of observations, the assimilation of observations to generate initial and reference conditions, model simulations, as well as post-processing of model data and the development of forecast products. For each application, benchmark datasets with up to 10 terabytes of data will be published online for training and machine learning tool-developments at the scale of the fastest supercomputers in the world. MAELSTROM machine learning solutions will serve as blueprint for a wide range of machine learning applications on supercomputers in the future.

2.2 Scope of this deliverable

2.2.1 Objectives of this deliverable

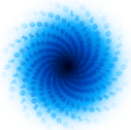
D4.2 Data Management Plan provides the initial outline of the data management plan including information on which data sets will be created in the project and how they will be made available. This document represents only the initial version where details may not be available yet, and it will be further developed over the course of the project.

2.2.2 Work performed in this deliverable

The work performed included, as per the DoA, the collection of the available descriptions of data sets to be produced by the project, through a questionnaire.

2.2.3 Deviations and counter measures

No deviations have been encountered.



3 Open Research Data Objectives

3.1 Open Research Data Pilot

As per the Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020[1], Research Data

> "Refers to information, in particular facts or numbers, collected to be examined and considered as a basis for reasoning, discussion, or calculation. In a research context, examples of data include statistics, results of experiments, measurements, observations resulting from fieldwork, survey results, interview recordings and images. The focus is on research data that is available in digital form."

The Open Research Data Pilot

> "aims to improve and maximise access to and re-use of research data generated by Horizon 2020 projects[2]"

and applies to data sets that are

> "needed to validate the results presented in scientific publications[2]".

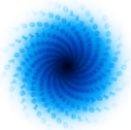
The Data Management Plan is expected to

> "specify what data will be open: detailing what data the project will generate, whether and how it will be exploited or made accessible for verification and re-use, and how it will be curated and preserved[2]".

3.2 MAELSTROM Research Data

As per the Description of Actions, MAELSTROM will generate and use model output from Weather & Climate (W&C) models, weather observations (including standard products but also crowed-sourced weather station data), data from energy production (e.g., wind farms or solar panels), as well as data from citizen sensing, crowd-sensing and tweeting sensors. Only de-personalised/anonymised data will be handled within MAELSTROM. The data will be unified via a data API that is organised by WP2 and mainly based on the CliMetLab tool[3]. The data will be used to generate benchmark datasets of MAELSTROM that will be published as open access data via S3 buckets. The datasets are made available in different tiers where possible, including a tier-1 version that can be used for training purposes that will not exceed a size that is not manageable by a standard laptop and larger tiers that can be used for scaling experiments with tens of terabytes for some of the applications. Since the data will be available as open source, verification and re-use will be guaranteed. The datasets will be curated by ECMWF as they are meant to become a valuable asset of the W&C community to allow

[1] http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf
[2] https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-dissemination_en.htm
[3] https://climetlab.readthedocs.io/en/latest/



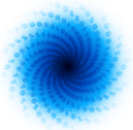
for quantitative evaluations of Machine Learning (ML) solutions. ECMWF will cover the cost for data curation.

3.3 Data Management Plan Questionnaire

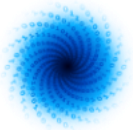
The following questionnaire has been provided to MAELSTROM technical work packages to gather the information for this first version of the Data Management Plan.

Table 1: MAELSTROM Data Management Plan Questionnaire

<Data set reference and name>	
Data set description	*Description of the data that will be generated or collected (or is already available to the project), its origin (in case it is collected), nature and scale and to whom it could be useful, and whether it underpins a scientific publication. Information on the existence (or not) of similar data and the possibilities for integration and reuse.* *Limitations?* *Constraints?*
Standards and metadata	*Reference to existing suitable standards of the discipline. If these do not exist, an outline on how and what metadata will be created.* *Will you generate proper metadata for you data?* *If yes: how do they look like?* *If no: why?* *Data format?* *Will there be a review process to quality- check the data?*
Data Sharing	*Description of how data will be shared, including access procedures, embargo periods (if any), outlines of technical mechanisms for dissemination and necessary software and other tools for enabling re-use, and definition of whether access will be widely open or restricted to specific groups. Identification of the repository where data will be stored, if already existing and identified, indicating in particular the type of repository (institutional, standard repository for the discipline, etc.).* *In case the dataset cannot be shared, the reasons for this should be mentioned (e.g. ethical, rules of personal data, intellectual property, commercial, privacy-related, security-related).* *License?* *Access URL?*
Archiving and preservation (including storage and backup)	*Description of the procedures that will be put in place for long-term preservation of the data. Indication of how long the data should be preserved, what is its approximated end volume, what the associated costs are and how these are planned to be covered.*



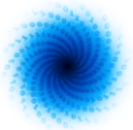
	At which Data Center do you want to store your data? *Is there an established workflow for your requested DOI process in place?* *According to which standards*



4 MAELSTROM Data Sets

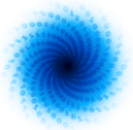
The following sections provide the responses by work packages.

maelstrom-yr	
Data set description	Contains gridded weather data for the Nordics. It contains both predictors (gridded weather forecasts) and predictand (gridded analysis fields). The forecasts are used operationally for the Nordic region on https://www.yr.no . The use of a subset of this dataset has been described in "Nipen, T. N., I. A. Seierstad, C. Lussana, J. Kristiansen, and Ø. Hov, 2020: Adopting Citizen Observations in Operational Weather Prediction. Bull. Amer. Meteor. Soc., 101, E43–E57." *Limitations* No *Constraints* No
Standards and metadata	The dataformat is NetCDF and the dataset follows the NetCDF/CF standard. The dataset contains metadata that describes the coordinate system, target and predictor variables, and their units. The metadata is stored in the NetCDF files. *Will you generate proper metadata for your data?* *If yes: how do they look like?* *If no: why?* *Data format* netCDF *Will there be a review process to quality-check the data?* The review process will consist of spot checks of the data performed by an independent person not involved with the data generation.
Data Sharing	Jupyter notebooks have been created to explore this dataset. CliMetLab is used to download for data loading. The Jupyter can be accessed through github and be used for public. *In case the dataset cannot be shared, the reasons for this should be mentioned (e.g. ethical, rules of personal data, intellectual property, commercial, privacy-related, security-related).* *License*



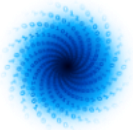
	The license of the data is described here: https://www.met.no/en/free-meteorological-data/Licensing-and-crediting *Access* https://github.com/metno/maelstrom-yr/blob/main/notebooks/demo_yr.ipynb
Archiving and preservation (including storage and backup)	*At which Data Center do you want to store your data?* ECMWF data cloud *Is there an established workflow for your requested DOI process in place? According to which standards* Not yet.

maelstrom-radiation	
Data set description	This is a dataset to train machine learning emulators for parametrisation schemes of the IFS weather forecast model. The data is collected from forecast simulations with the IFS. The data has a size of several terabytes and will be useful to researchers and developers of weather and climate models. Emulators that are trained from the Deliverable can only be as good as the original scheme, unless the data is used as inputs to generate more sophisticated output fields. This data is for learning the emulation of the ECMWF radiation scheme, TripleClouds, found in the ecRad package (https://github.com/ecmwf/ecrad). Building an accurate emulator of radiative heating could accelerate weather and climate models partially by enabling the use of GPU hardware within our models. *Limitations* *No* *Constraints* *No*
Standards and metadata	Within the netCDF files the data follows CF conventions. For speed of learning this metadata is removed in the packing in TFRecords format, however this data is a repackaging of the netCDF which can be consulted for referencing metadata. *Will you generate proper metadata for you data?* *If yes: how do they look like?* *If no: why?* *Data format?* netCDF and TFRecords



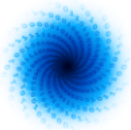
	Will there be a review process to quality- check the data? Yes
Data Sharing	Jupyter notebooks have been created to explore this dataset. CliMetLab is used to download for data loading. The Jupyter can be accessed through github and be used for the public. *License* https://storage.ecmwf.europeanweather.cloud/MAELSTROM_AP3/LICENCE.txt *Access URL* https://git.ecmwf.int/projects/MLFET/repos/maelstrom-radiation/browse/notebooks/demo_radiation.ipynb
Archiving and preservation (including storage and backup)	The approximated end volume will be ~ 10 TB *At which Data Center do you want to store your data?* ECMWF cloud storage *Is there an established workflow for your requested DOI process in place? According to which standards* Not yet, but several options to obtain a DOI are investigated.

maelstrom-nogwd	
Data set description	Contains the input/output dataset for learning non-orographic gravity wave drag, as described in https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2021MS002477. Data is grouped by forecast start date. Data has been pre-processed into inputs "x" and outputs "y". "x" contains vertical profiles of winds & temperature plus surface values of pressure and geopotential. "y" contains the wind increments due to parametrised non-orographic gravity wave drag. The machine learning task is to predict y given x. Unlike many ML tasks within the field of weather and climate, this task can be predicted independently for each column of the atmosphere. *Limitations* No *Constraints* No

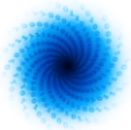


Standards and metadata	*Will you generate proper metadata for you data?* *If yes: how do they look like?* *If no: why?* No, because the data is only available in a friendly format for machine learning, with the variables concatenated together for reading speed. This makes adding metadata challenging. A full description of the variables and their ordering is provided in the data documentation. *Data format* netCDF *Will there be a review process to quality- check the data?* Yes
Data Sharing	*License* https://storage.ecmwf.europeanweather.cloud/NOGWD_L91_PUBLIC/LICENCE.txt *Access URL* https://git.ecmwf.int/projects/MLFET/repos/maelstrom-nogwd/browse/notebooks/demo_nogwd.ipynb
Archiving and preservation (including storage and backup)	The approximated end volume will be ~3Gb *At which Data Center do you want to store your data?* ECMWF cloud storage *Is there an established workflow for your requested DOI process in place? According to which standards* Not yet

maelstrom-downscaling

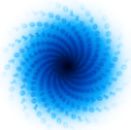


Data set description	The two datasets will comprise various meteorological quantities which are considered to convey relevant information for downscaling the 2m temperature. In addition to the 2m temperature itself, these are temporally variable quantities such as the surface enthalpy fluxes, the cloud cover, or the temperature at different pressure levels, but also constant features such as the surface elevation. The data will be obtained from historical IFS HRES-forecasts (from 2016 onwards) in the short-range (max. lead time 24 hours). This provides data is used at 0.1 resolution on a spherical grid. Central Europe is chosen as the target region since the topography is highly complex in this area. As part of deliverable D1.1., a subset of the data (2m temperature and surface elevation at initial times of the IFS model) are used and artificially coarsened onto a 0.8° grid to imitate the input data for a supervised downscaling application. In the future, the application will approach a real-case application by e.g. applying ERA5 reanalysis data (on a coarser 0.3°-grid) as input data. Additionally, more quantities will reside in the dataset. In the future, other model data sources will be considered such as model forecast from the multiscale ICON-model operated at DWD (German Weather Service). The latter would enable supervised training from a coarse-grained global grid at 13 km resolution to a high-resolved grid at 2.2 km resolution over Germany. *Limitations* No *Constraints* No
Standards and metadata	*Will you generate proper metadata for your data?* *If yes: how do they look like?* Model data are commonly shipped along with metadata that typically follows the so-called CF-conventions. Thus, information on the data source, the underlying grid, the validity time (initial time of the model and the lead time) etc. are already part of the processed data and no further metadata creation process is required. However, pre-processing operations using CDO will also be tracked. *If no: why?* *Data format* netCDF *Will there be a review process to quality- check the data?* Yes



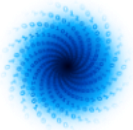
Data Sharing	Data sharing of the IFS HRES data is possible in scope of the project. Namely, deliverable D1.1. already realized a publication of some IFS HRES data under the Apache 2.0 License. The IFS HRES data itself was processed using the Meteorological Archival and Retrieval System (MARS). The ERA5 reanalysis data can be obtained from the C3S climate data store. The associated license can be viewed here: https://apps.ecmwf.int/datasets/licences/copernicus/ The ICON model data is distributed via DWD's Open Data Server. Terms and conditions are claimed here. *License* Yes *Access URL* *https://git.ecmwf.int/projects/MLFET/repos/maelstrom-downscaling-ap5/browse*
Archiving and preservation (including storage and backup)	*At which Data Center do you want to store your data?* The data is stored in the ECMWF cloud storage. *Is there an established workflow for your requested DOI process in place? According to which standards* Not yet.

maelstrom-power-production	
Data set description	A dataset plugin for CliMetLab for the dataset climetlab-plugin-a6/maelstrom-production-forecasts. It consisted of Weather forecast data and power production data and described in Deliverable D1.1. *Limitations* No *Constraints* No



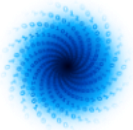
Standards and metadata	*Will you generate proper metadata for you data?* Not needed, all metadata are included. *Data format* netCDF *Will there be a review process to quality- check the data?* Yes
Data Sharing	Jupyter notebooks have been created to explore this dataset. CliMetLab is used for data loading. The Jupyter can be accessed through github and be used for public. *License* Apache 2.0 *Access URL* https://github.com/4castRenewables/climetlab-plugin-a6/blob/main/notebooks/demo_maelstrom_production_forecasts.ipynb
Archiving and preservation (including storage and backup)	The approximated volume will be ~ GB - ~TB *At which Data Center do you want to store your data?* ECMWF cloud storage *Is there an established workflow for your requested DOI process in place? According to which standards* Not yet.

Social media data	
Data set description	The data are obtained by the Twitter API. At the moment, the access is not granted, application for data access are still ongoing at Twitter.
Standards and metadata	The data are stored as obtained from Twitter through the API. We will eventually store the data in a database and dump the data as csv to the ECMWF S3 bucket. *Will you generate proper metadata for you data?* *If yes:* we will use the incoming policy from twitter, e.g.location information, if exists. Further details are not yet clear

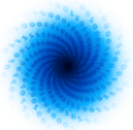


	Data format? Csv *Will there be a review process to quality- check the data?* No
Data Sharing	We will build a database and dump data to the S3 bucket. Sharing is not allowed due to the Twitter policies. Only if Twitter has agreed, we can share. *License?* None *Access URL?* None
Archiving and preservation (including storage and backup)	We follow the Twitter rules, cf. https://developer.twitter.com/en/developer-terms/commercial-terms

maelstrom-ens10	
Data set description	ENS10 is an open dataset developed as part of MAELSTROM to enable research into machine learning of weather ensemble prediction systems via post-processing. It consists of the model output data of ECMWF "hindcast" experiments. These are global ensemble forecasts with 10 ensemble members that are spread over 20 years (1998-2017) with two forecasts per week. The dataset holds 10 ensemble members on 0.5 degree latitude/longitude grids. ENS10 weighs 3TB, thus for smaller-scale experiments we provide a smaller, tier-1 version called ENS5-mini, which weighs 9GB and consists of 5 members, 10 years, one day into the future, and is cropped over central Europe. One limitation of the dataset is that it does not include the full, 51-member ensemble used in production systems, nor all fields used for weather prediction. Extending ENS10 will be researched as part of this project.
Standards and metadata	The metadata of ENS10 is included within its data format – GRIB – which contains spatial information and units of each of the provided fields. Since the data consists of hindcasts from the ECMWF system, the information contained with it is sound. Therefore, our review process of the data consists of data-scientific inspection of the dataset and testing it on machine learning models.
Data Sharing	The dataset, in its raw GRIB format, is already uploaded onto the European Weather Cloud S3 servers and mirrored on a public ETH Zurich storage server. Both links are public and available as part of the ECMWF website, ETH Zurich, and the CliMetLab plugin we provide to accompany the dataset. The dataset is licenced under the ECMWF general data access licence (https://apps.ecmwf.int/datasets/licences/general/, CC BY 4.0). Access links: • Via CliMetLab: climetlab-maelstrom-ens10



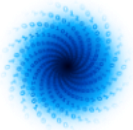
	• Via the European Weather Cloud: https://storage.ecmwf.europeanweather.cloud/MAELSTROM_AP4 • Via the ETH Zurich mirror: http://spclstorage.inf.ethz.ch/projects/deep-weather/ENS10/ , http://spclstorage.inf.ethz.ch/projects/deep-weather/ENS5mini/
Archiving and preservation (including storage and backup)	There is not yet a DOI for the dataset until its finalization w.r.t. fields and ensemble member counts. ENS10 will follow the European Weather Cloud data center policy for long-term retention. The mirror at ETH Zurich is kept as part of the Scalable Parallel Computing Laboratory servers, which contain more datasets, and will stay on for as long as the lab is operational, with no additional storage costs.



5 Conclusion

This initial Data Management Plan has identified a number of data sets for each of the technical work packages, identifying the required details (where possible) on what data will be open, how it will be made accessible, and how it will be curated.

The Data Management Plan is to be seen as a living document and will be reviewed and revised periodically to ensure that information contained therein is up-to-date and correct.



Document History

Version	Author(s)	Date	Changes
0.1	Daniel Thiemert (ECMWF)	20/09/2021	Initial version
1.0	Daniel Thiemert (ECMWF)	21/09/2021	Final Version

Internal Review History

Internal Reviewers	Date	Comments
Peter Dueben (ECMWF)	21/09/2022	Approved with comments

Estimated Effort Contribution per Partner

Partner	Effort
ECMWF	0.2
Total	**0.2**