MAchinE Learning for Scalable meTeoROlogy and climate

The 2nd Bootcamp of MAELSTROM

Ana Prieto Nemesio

Jan Mirus

www.maelstrom-eurohpc.eu

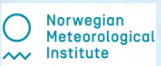
D4.6

The 2nd MAELSTROM Bootcamp

Author(s): Ana Prieto Nemesio
Jan Mirus

Dissemination Level: Confidential

Date: 24/11/2023
Version: 1.0
Contractual Delivery Date: 24/11/2023
Work Package/ Task: WP4/ T4.6
Document Owner: ECMWF
Contributor(s): Peter Dueben
Status: Final



MAELSTROM

Machine Learning for Scalable Meteorology and Climate

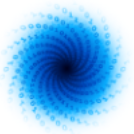
Research and Innovation Action (RIA)
H2020-JTI-EuroHPC-2019-1: Towards Extreme Scale Technologies and Applications

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

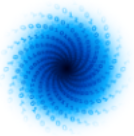
The MAELSTROM project has received funding from the European High-Performance Computing Joint Undertaking (JU) under grant agreement No 955513. The JU receives support from the European Union's Horizon 2020 research and innovation programme and United Kingdom, Germany, Italy, Luxembourg, Switzerland,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



Contents

Figures

Tables



1 Executive Summary

MAELSTROM's second Bootcamp was hosted by ECMWF on its Reading, UK site between November 8th and November 10th, 2023.

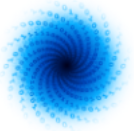
The aim of this three-day Boot camp was to provide participants with an overview of the MAELSTROM Applications of machine learning (ML) in weather and climate on HPC systems. Following introductory talks on the scientific foundations of meteorology, ML, and HPC, participants were given the opportunity to join a more detailed introduction to machine learning, and in particular deep learning, or to follow a tutorial into performance modelling of machine learning applications for more advanced participants. Afterwards, participants were joining smaller groups to delve into exploration of the ML applications developed within the MAELSTROM Project.

The Bootcamp covered the following topics:

1. Overview of the MAELSTROM Project.
2. Training participants about ML methods, meteorology and HPC systems.
   Training material included:
   a. Sessions of interactive introductory-level lectures on the above topics.
   b. Lectures about the MAELSTROM applications, focusing on the different ML approaches implemented for each of the use cases.
   c. Hands-on tutorials for participants to interactively follow across the boot camp.

Due to travelling visa complications, in this 2nd Bootcamp we could not have a representative for Application 4 "Improved ensemble predictions in forecast post-processing". However, all other applications were represented and supervised by a team of 1-3 advisors per application.

The Deliverable also provides information on the results of a User Survey that was performed in the same week as the MAELSTROM Dissemination Workshop and the Bootcamp.



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.

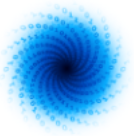
The MAELSTROM machine learning applications will cover all important components of the workflow of weather and climate predictions including the processing of observations, the assimilation of observations to generate initial and reference conditions, model simulations, as well as post-processing of 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 a 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

D4.6, the second MAELSTROM Bootcamp, aimed to:

- Train participants on the MAELSTROM applications so they can have a better understanding on how to build such datasets and models, as well as comprehend the opportunities and challenges behind the use of ML for weather and climate.
- Bring together young scientists from the fields of weather and climate science, ML and HPC to learn from each other and to generate synergies between the scientific domains.
- Gather user feedback about the usage of MAELSTROM applications and datasets, as well as the experience across the workshop.



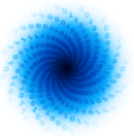
- Provide an overview of the scope behind MAELSTROM project and the work done across the applications

2.2.2 Work performed in this deliverable

Deliverable 4.6 provides a summary of the execution of the 2nd MAELSTROM Bootcamp. This second Bootcamp continues the work already started during the celebration of the first workshop in 2022 to train the next generation of ML scientists in the domain of weather and climate science. As the concept to train the participants hands-on with the MAELSTROM applications has worked out very well during the first Bootcamp, the concept was repeated during this event. Details regarding the planning, agenda, and training curriculum are described in Section 3. This Deliverable also presents results gathered from the user survey circulated via the MAELSTROM webpage, social media and the participants of the MAELSTROM events. The results are summarised and described in Section 4.

2.2.3 Deviations and counter measures

There are no significant deviations from the planned contributions of the deliverable.



3 MAELSTROM Boot Camp Outline

3.1 The agenda of Bootcamp

The 2nd MAELSTROM Bootcamp was held at ECMWF Reading facilities from the 8th to the 10th of November, 2023. The exact programme can be found on the ECMWF events website[1] or the MAELSTROM webpage[2]. The Bootcamp spanned three days: on Day 1, we focused on providing participants with an introduction to the MAELSTROM Project, ML applied to weather and climate science, and HPC systems. Afternoon sessions included sessions to gather a better understanding of deep learning and benchmarking model performance. The first day concluded with an overview of the MAELSTROM applications provided by each of the application developers so that participants could choose between them.

On Day 2, participants continued working on the different hands-on training materials about the use of ML benchmark datasets and solutions provided by each MAELSTROM application. Before lunch, all application teams gathered together to give a short pitch about their first insights and experience working on each of the applications.

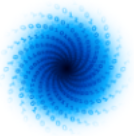
On the 3rd day, participants worked during the morning and wrapped up their work to conclude the Bootcamp presenting their results in short 10-15 minute presentations to the supervisors and participants. They summarised the results trying to focus on what they have learned during the training, main challenges, and potential avenues to solve those, and were answering questions by the audience.

Wednesday, 8 November 2023		
	Morning session	
09:00 to 09:30	A high-level introduction into MAELSTROM, Earth sciences, machine learning, and high performance computing	Peter Düben (ECMWF)
09:30 to 10:30	A closer look into high performance computing	Andreas Herten (JSC)
10:30 to 11:00	Coffee break	
11:00 to 12:00	A closer look into machine learning for weather and climate applications	Peter Düben (ECMWF)
12:00 to 12:30	Q&A	
12:30 to 13:30	Lunch break	
	Afternoon session	
	Group photo	
13:30 to 14:30	Group A: A demo of performance models	Karthick Panner Selvam (University of Luxembourg)
13:30 to 14:30	Group B: A shallow entry into deep learning and other machine learning methods	Peter Düben (ECMWF) Ana Prieto Nemesio (ECMWF)
14:30 to 15:00	Organisation into groups	
15:00 to 18:00	Free work in groups Coffee / tea available at 15:30	
19:00 to 21:00	**Boot Camp pub dinner**	
Thursday, 9 November 2023		
	Morning session	
09:00 to 09:30	Progress update of groups	
09:30 to 11:00	Free work in groups	
11:00 to 11:30	Coffee break	
11:30 to 12:30	Free work in groups	
12:30 to 13:30	Lunch break	
	Afternoon session	
13:30 to 15:30	Free work in groups	
15:30 to 16:00	Coffee break	
16:00 to 17:30	Free work in groups	
17:30 to 18:00	Progress update of groups	
Friday, 10 November 2023		
	Morning session	
09:00 to 11:00	Free work in groups Including coffee break at 10:30	
11:00 to 13:00	Summary of results and close	

3.2 Training curriculum

To train participants on the use of machine learning for weather and climate application on HPC systems, the training curriculum of this second Bootcamp included both introductory sessions and hands-on training.

[1] https://events.ecmwf.int/event/351/
[2] https://www.maelstrom-eurohpc.eu/article?topic=bootcamp-2023

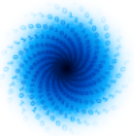


Introductory lectures were scheduled on the first day to ensure participants could have a good overview of the theoretical background of (scalable) Machine Learning, Weather and Climate domain, and HPC systems. The lectures include:

- MAELSTROM introduction and overview (Peter Dueben)
- Introduction to Machine Learning in weather and climate (Peter Dueben)
- A closer look into HPC (Andreas Herten)
- Introduction to Deep Learning (Ana Prieto Nemesio)
- Introduction to Performance Predictive Model (Karthick Panner Selvam)

Application developers curated lectures and hands-on materials to teach the participants about their applications. Application developers designed the training materials including the lectures and exercises. Exercises were implemented using Jupyter Notebooks to facilitate collaboration and interactivity between participants and tutors. All exercises and hands-on materials were executed using Jupyter-JSC[3], a customised JupyterHub provided as part of Jülich Supercomputing Centre (JSC) to facilitate the use of the HPC ecosystem using a web-based interactive development environment for Jupyter notebooks, code, and data. Ahead of the boot camp participants were given instructions about Jupyter-JSC as well as login credentials. The materials were published and can be accessed through Gitlab[4]. The slides of presentations were shared with the participants via gitlab as well.

[3] https://jupyter-jsc.fz-juelich.de/hub/login?next=%2Fhub%2Fhome
[4] https://gitlab.jsc.fz-juelich.de/esde/training/maelstrom_bootcamp/-/tree/master/BootCamp_2023-11



4 User Survey and Feedback

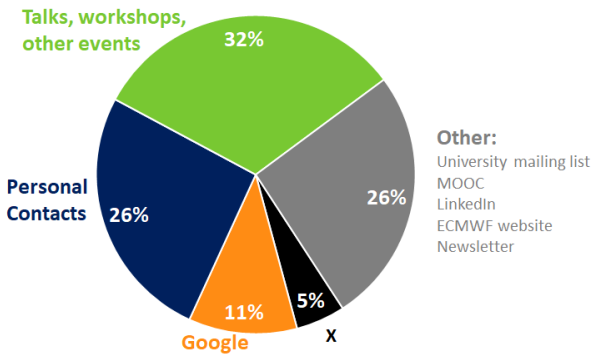
During the week of the Dissemination Workshop and the Boot Camp, we also performed a short User Survey that was advertised via the MAELSTROM webpage, social media, and at the MAELSTROM events. The aim of the user survey was to gather feedback about the interest and usability of the different MAELSTROM applications.

The questions of the user survey were:

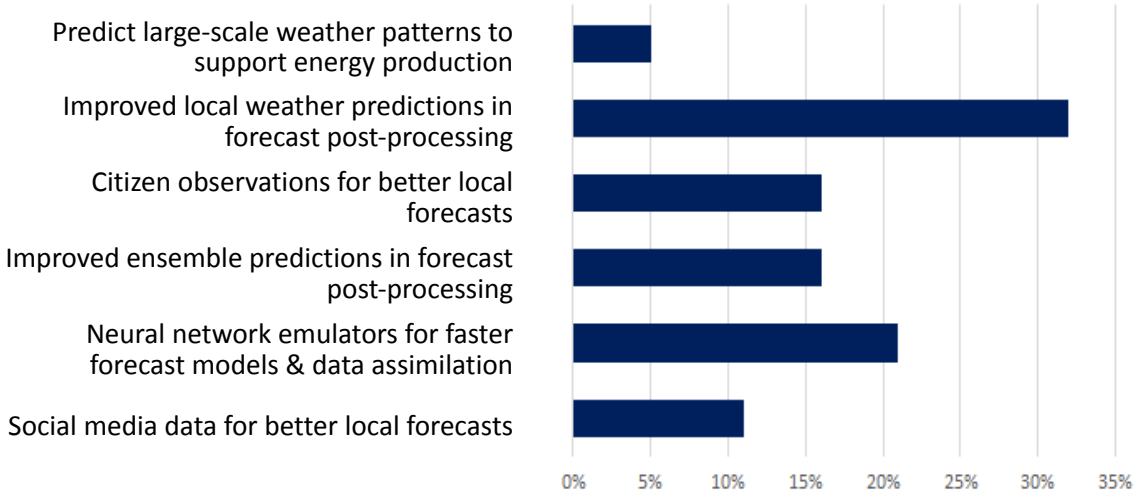
- How did you become aware of project MAELSTROM in the first place?
  (Twitter a.k.a. X | Google | Personal contacts | Scientific papers | Talks, workshops or other events | Other: … One answer allowed)
- Which MAELSTROM applications have you used?
  (Predict large-scale weather patterns to support energy production | Improved local weather predictions in forecast post-processing | Citizen observations for better local forecasts | Improved ensemble predictions in forecast post-processing | Neural network emulators for faster forecast models & data assimilation | Social media data for better local forecasts. Multiple answers allowed)
- Which MAELSTROM datasets have you used?
  (Datasets for 2m temp. and precipitation short-range forecasts | Dataset to emulate radiation | Dataset for ensemble predictions | Dataset for 2m temperature downscaling | Dataset for energy production forecast. Multiple answers allowed)
- Since you said you have used our datasets: Did you find it easy to download the data?
  (5-point scale)
- Did you find the Jupyter notebooks useful and easy to use?
  (5-point scale)
- What have you used the MAELSTROM datasets for?
  (Research | Teaching | Benchmarks | Other: …. Multiple answers allowed)
- What have you used the MAELSTROM notebooks for?
  (Research | Teaching | Benchmarks | Other: …. Multiple answers allowed)
- Did you find the documentation on GitHub useful?
  (5-point scale)
- What did you miss when using the datasets and notebooks?
  (open-ended question)
- Would you consider using CliMetLab to create and share a dataset?
  (yes, because: … / no, because: … / What is CliMetLab? One answer allowed.)
- Do you have any other comments for us?
  (open-ended question)

We yielded 19 completed questionnaires, which doesn't allow for advanced statistical analyses, but represents a good response rate of more than 50%, proving a motivated and committed audience.

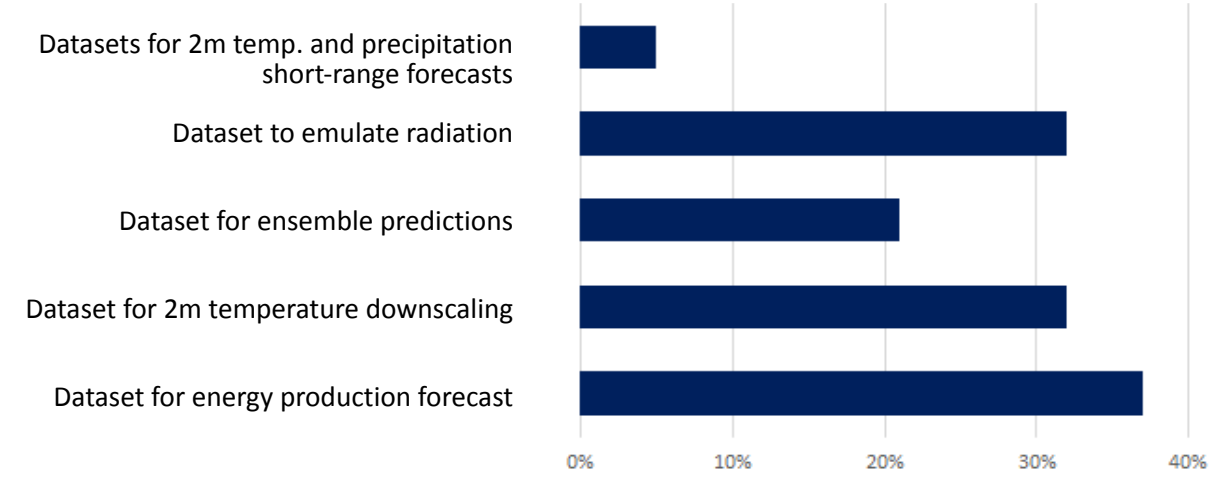
How did you become aware of project MAELSTROM in the first place?



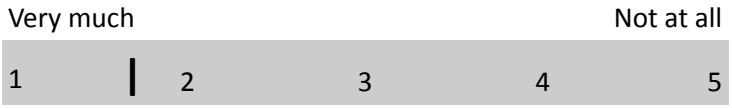
Which MAELSTROM applications have you used?



Which MAELSTROM datasets have you used?

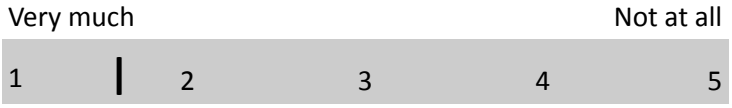


Did you find it easy to download the data?



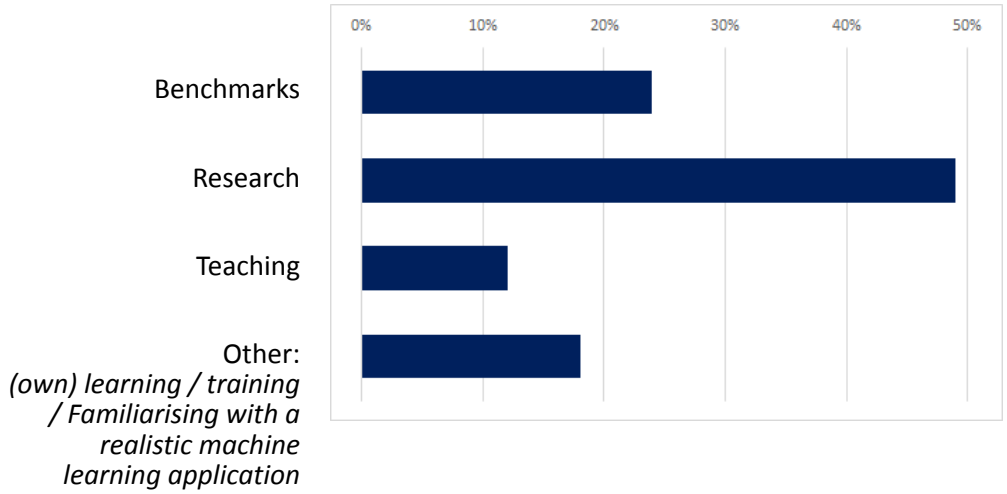
mean: 1.7

Did you find the Jupyter notebooks useful and easy to use?

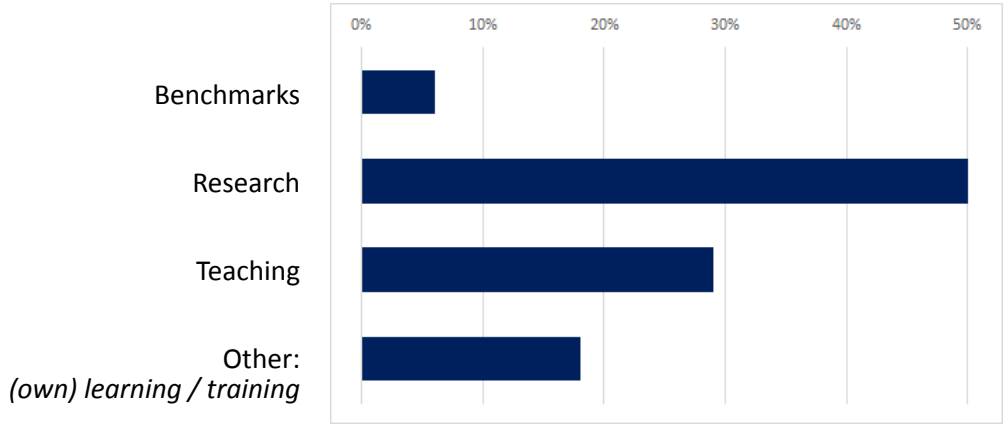


mean: 1.6

What have you used the MAELSTROM datasets for?



What have you used the MAELSTROM notebooks for?



Did you find the documentation on GitHub useful?

Very much — Not at all

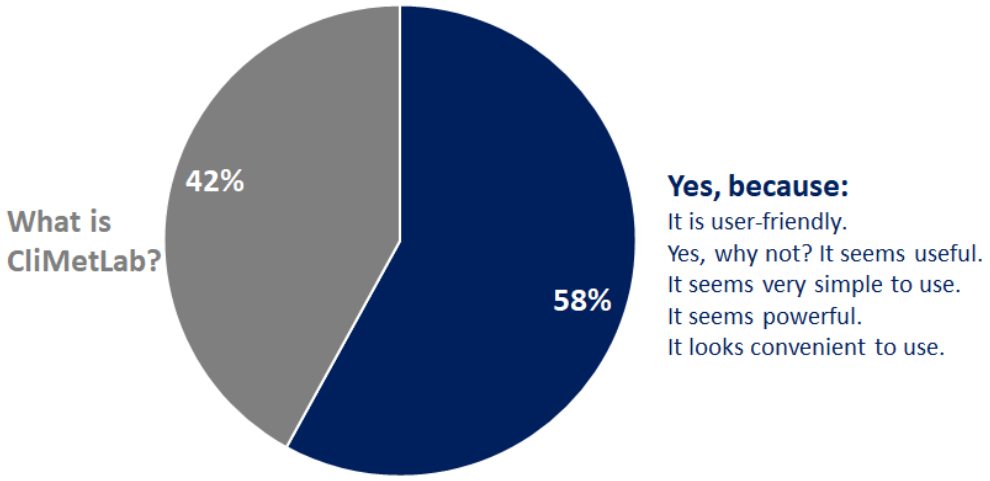
1 | 2 | 3 | 4 | 5

mean: 1.7

What did you miss when using the datasets and notebooks?

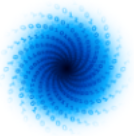
Maybe more reasoning on why using some NN architectures and not others.
Data preprocessing.
There were some issues with the notebooks, for example we were not able to output the loss of the machine learning model.

Would you consider using CliMetLab to create and share a dataset?



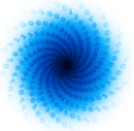
Do you have any other comments for us?

Thanks for the bootcamp! Enjoyed
A longer boot camp
Very helpful workshop!
Thanks for your valuable efforts.
The Boot Camp was interesting to learn about the application of machine learning. However, the duration is too short to understand how the model works.
Thanks for your efforts!



We interpret the survey outcome as follows:

Putting datasets and Jupyter notebooks at free disposal was and is a suitable means to disseminate the outcome of project work, as it provides value to individuals of the scientific community, research and own learning being the chief use case.
We managed to keep the threshold low, as easy download, notebook handling and GitHub support was widely acknowledged.
If we were to plan another bootcamp: It should not be shorter, but rather longer and with more depth, to quench the thirst for understanding of the participants, who arrive with quite different levels of knowledge.



5 Participants, tutors and networking

From Wednesday to Friday, we held the MAELSTROM Bootcamp, with 24 external participants on site who were learning more about how to work with machine learning in general, and the MAELSTROM applications and datasets in particular. Training about the use of the MAELSTROM applications was provided by 10 tutors across MAELSTROM partners, including:

- Thomas Nipen (MetNor)
- Fabian Emmerich, Kristian Ehlert, Oliver Kindler (4cast)
- Matthew Chantry, Ana Prieto Nemesio, Peter Dueben (ECMWF)
- Michael Langguth, Andreas Herten, Stepan Nassyr (JSC)

To foster networking across the duration of the Boot Camp, we had coffee breaks and joint dinner activities where participants and tutor had opportunities to connect and further discuss topics among each other. At the end of the Bootcamp, the option to voluntarily share contact details between participants and tutors was facilitated.

Figure 1: Group photo of the participants of the MAELSTROM Bootcamp at ECMWF Weather Room.

Figure 2: At work in the ECMWF council chamber.

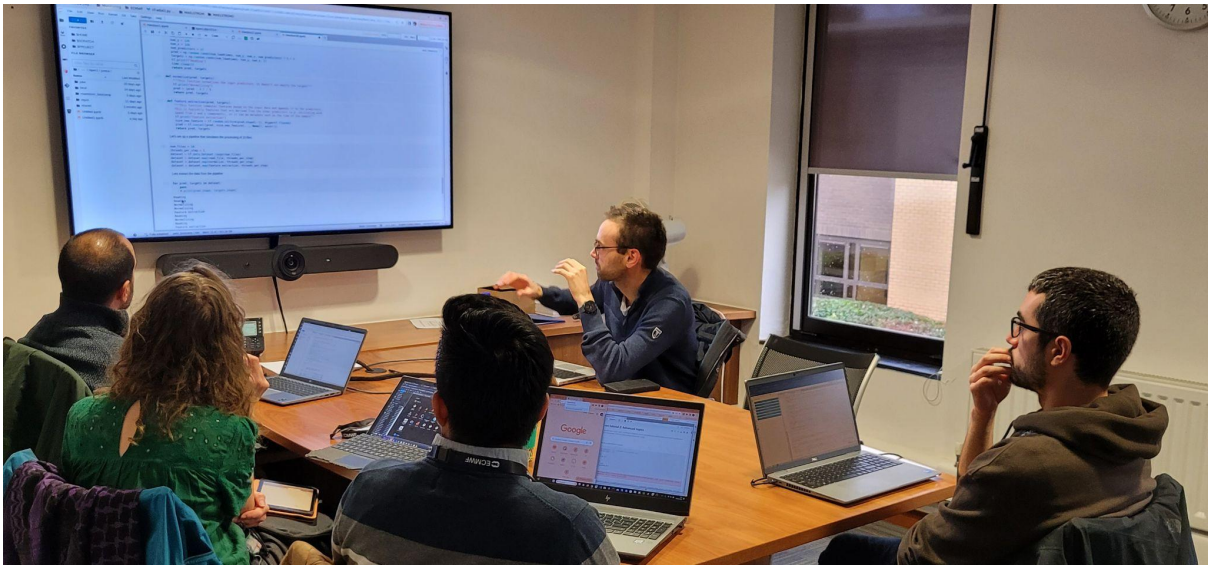
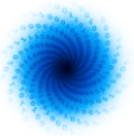


Figure 3: Intense training sessions in small groups.

Figure 4: Networking dinner after a day of hard work



Document History

Version	Author(s)	Date	Changes
1.0	Jan Mirus (4cast)	23/11/2023	Initial content
1.1	Ana Prieto Nemesio (ECMWF)	26/11/2023	Main lot of content added
1.2	Peter Dueben	27/11/2023	Additional content, remarks, corrections
1.3	Jan Mirus	28/11/2023	Survey results & images added

Internal Review History

Internal Reviewers	Date	Comments
Name (Organisation)	dd/mm/yyyy	
Thomas Nipen (Norwegian Meteorological Institute)	04/12/2023	Accepted with minor revisions
Mats Brorsson (University of Luxembourg)	04/12/2023	Accepted with minor revisions

Estimated Effort Contribution per Partner

Partner	Effort
ECMWF	0.2PM
4cast	0.2PM
Total	**0.4PM**