europlanet 2024
Research Infrastructure

H2020-INFRAIA-2019-1

Europlanet 2024 RI has received funding from the
European Union's Horizon 2020 Research and Innovation Programme under

Grant agreement no: 871149

Deliverable D10.7

Deliverable Title:	ML 3rd Year Report
Due date of deliverable:	31/01/2023
Nature[1]:	R
Dissemination level[2]:	PU
Work package:	10
Lead beneficiary:	INAF
Contributing beneficiaries:	IWF-OeAW, KNOW, UNIPASSAU, ACRI-ST, DLR, IAP-CAS, AOP
Document status:	Final

Start date of project:	01 February 2020
Project Duration:	54 months
Co-ordinator:	Prof Nigel Mason, University of Kent

1. **Nature:** D
2. **Dissemination level:**

PU	PP	RE	CO
Public	Restricted to other programme participants (including the Commission Service)	Restricted to a group specified by the consortium (including the Commission Services)	Confidential, only for members of the consortium (excluding the Commission Services)

Executive Summary / Abstract:

This annual report summarizes the work done in Work Package 10 'Machine Learning Solutions for Data Analysis and Exploitation in Planetary Sciences' during the third year of Europlanet 2024 Research Infrastructure. The main aims of the work package are to foster wider use of machine learning technologies in data driven space research and to provide open-source machine learning code developed for specific science cases. Work Package 10 is organized around six tasks that target management and coordination of the activities, the development of machine learning based data analysis code and the dissemination of the tools, as well as integration of the results into VESPA, GMAP and SPIDER where appropriate. Despite delays in the development work due to the ongoing COVID-19 pandemic, work on all six tasks has been progressing. Developments of six science cases are finished and work on the remaining science cases is progressing. We conducted three workshops at the EPSC 2022 conference, introducing four of our machine learning pipelines. We put up tutorials on our machine learning portal as well as on our public GitHub repositories. We also started to compile a Jupyter book about the content created in our work package. ML organized machine learning sessions at EGU22 and EPSC2022, and had presentations at many conferences (EGU22, EPSC2022, ECML PKDD 2022, ESWW 2022, AGU Fall Meeting 2022). We have collaborations with national (FWF project at IWF) as well as international (EU Horizon 2020 EXPLORE project) research projects, and started a series of fireballs workshops together with NA2. An EPN-TAP server was set up at IWF, on which we started to integrate first data sets of our science cases into VESPA. Furthermore, next steps were undertaken to include our pipelines in SPIDER.

Table of Abbreviation	
AGU	American Geophysical Union

ASPP	Atrous Spatial Pyramidal Pooling
BS	Bow Shock
CIR	Corotating Interaction Region
CME	Coronal Mass Ejection
CNN	Convolutional Neural Network
D	Deliverable
DMP	Data Management Plan
DTM	Digital Terrain Model
EGI	European Grid Infrastructure
EGU	European Geophysical Union
EOSC	European Open Science Cloud
EPN-2024-RI	Europlanet 2024 Research Infrastructure
EPSC	Europlanet Science Congress
ESA	European Space Agency
ESWW	European Space Weather Week
GAN	Generative Adversarial Network
GMAP	Geologic MApping of Planetary bodies
GPU	Graphics Processing Unit
HCS	Heliospheric Current Sheet
ICA	Independent Component Analysis
ICME	Interplanetary Coronal Mass Ejection
IMF	Interplanetary Magnetic Field
IoU	Intersection over Union
JpGU	Japan Geoscience Union
JRA	Joint Research Activity
LPSC	Lunar and Planetary Science Conference
LSTM	Long-Short-Term Memory Network
MASCS	Mercury Atmospheric and Surface Composition Spectrometer
MESSENGER	MErcury Surface, Space ENvironment, GEochemistry, and Ranging
ML	Machine Learning
MP	Magnetopause

MS	Milestone
NA	Networking Activity
PMC	Project Management Committee
SDA	Scientific Data Application
SPIDER	Sun Planet Interactions Digital Environment on Request
TRL	Technology Readiness Level
UMAP	Uniform Manifold Approximation and Projection
VA	Virtual Access
VESPA	Virtual European Solar and Planetary Access
WP	Work Package

1. Explanation of WP10 Work & Overview of Progress

a. Objectives and Description of Work

The objectives and description of work for Work Package (WP) 10 'JRA4 ML - Machine Learning Solutions for Data Analysis and Exploitation in Planetary Sciences' are as follows, quoted from the proposal:

> JRA4 will develop Machine Learning (ML) powered data analysis and exploitation tools optimised for planetary science and integrate expert knowledge on ML into the planetary community. All tools will also be linked via the VA services of VESPA, GMAP and SPIDER (where appropriate).
>
> The main objectives are:
>
> - to develop ML tools, designed for and tested on planetary science cases submitted by the community, and to provide sustainable, open access to the resulting products, together with support documentation
> - to foster wider use of ML technologies in data driven space research, demonstrate ML capabilities and generate a wider discussion on further possible applications of ML
> - to identify scientific and commercial applications for the ML tools developed through the JRA tasks
>
> **Description of work**
>
> This JRA will be led by IWF-OEAW, co-led by KNOW, and organised around 6 tasks. It will develop ML powered data analysis and exploitation tools that target a set of representative scientific cases selected from about a dozen proposals for specific applications of ML in planetary science submitted by the scientific user community in the course of proposal preparation. Software developed during the JRA will be open source (Apache License 2.0), thoroughly documented and

available via a git service, so that all results can be used freely, and further developed and extended by the community.

Work Package Beneficiaries

Apart from the WP lead, IWF-OEAW, there are eight beneficiaries contributing to our WP. Table 1 lists the acronyms of the WP beneficiaries as used in the Europlanet 2024 Research Infrastructure (EPN2024-RI) proposal and their corresponding institutions.

Due to sanctions against Russia, the participation of LMSU has been terminated. One science case about the automatic detection of boundary crossings around the planet Mercury was proposed by LMSU. Since this science case was finished before the sanctions against Russia were installed, the termination of LMSU's participation in Europlanet 2024 Research Infrastructure does not have any effect on WP10.

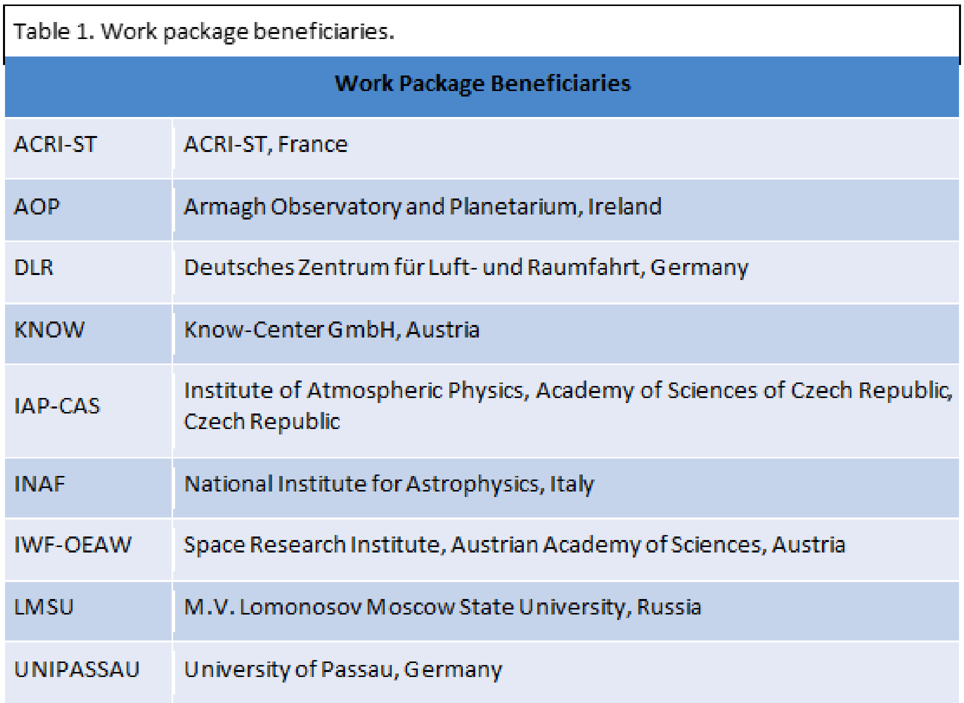


Table 1. Work package beneficiaries.

Work Package Beneficiaries	
ACRI-ST	ACRI-ST, France
AOP	Armagh Observatory and Planetarium, Ireland
DLR	Deutsches Zentrum für Luft- und Raumfahrt, Germany
KNOW	Know-Center GmbH, Austria
IAP-CAS	Institute of Atmospheric Physics, Academy of Sciences of Czech Republic, Czech Republic
INAF	National Institute for Astrophysics, Italy
IWF-OEAW	Space Research Institute, Austrian Academy of Sciences, Austria
LMSU	M.V. Lomonosov Moscow State University, Russia
UNIPASSAU	University of Passau, Germany

Science Cases

The science cases proposed by the planetary science community during proposal preparation are listed in Table 2. The proposal by GMAP covers different cases dealing with the detection and classification of various planetary surface features, as for example mounds and pits.

Table 2: list of science cases

Proposer	Science Case
IAP-CAS	Detection of plasma boundary crossings at planetary magnetospheres and solar wind
	Classification of plasma wave emissions in electromagnetic spectra

INAF	Mineral identification via reflectance spectra [possible applications foreseen in GMAP]
DLR	Classification of surface composition on the surface of Mercury [resulting data products can be used for GMAP]
AOP	Abundance of asteroids in Earth-like orbits from STEREO images
GMAP	Automatic recognition and analysis of planetary surface features
IWF-OEAW	Detection and classification of CMEs and CIRs in in-situ solar wind data
LMSU	Search for magnetopause/shockwave crossings on Mercury based on MESSENGER data

The science case by AOP needed to be re-formulated, since it was not do-able the way it was proposed. More details about the new science case can be found later in the according section.

Deliverables and Milestones

There are nine deliverables and three milestones for our WP, listed in Table 3.

Table 3: list of deliverables (D) and milestones (MS)

Abbreviations	Description	Month due	Finished
D10.1	Annual Report 1	M12	✓
D10.2	Annual Report 2	M24	✓
D10.3	Tutorial on Machine Learning and Basic How Tos (initial release)	M31	✓
D10.4	Demonstrator and Documentation of Data-Processing Techniques	M42	
D10.5	Demonstrator and Documentation of Time-based Signal Analysis and Automatic Classification Tool	M42	
D10.6	Demonstrator and Documentation of General Classification Toolset	M42	
D10.7	Annual Report 3	M36	
D10.8	Tutorial on Machine Learning and Basic How Tos (final release)	M42	
D10.9	Annual Report 4	M48	
MS11	Requirements for ML tools documented	M4	✓
MS51	ML Demonstrators implemented and tested	M30	✓

MS86	ML Demonstrators fully validated and integrated	M42	

b. Explanation of the work carried in WP

Task 1 - Management and Coordination

This task oversees the management of ML JRA4, coordinates the activities within the WP and with the other WPs and reports to the PMC.

Task 2 - Requirements for Machine Learning, Tool Validation and Communication

Infrastructure

There is now more information about our activity on the ML Portal, e.g., more information about the science cases, presentations, news regarding ML conferences, sessions and tutorials.

We put the draft version of a Jupyter book on our GitHub repository (https://github.com/epn-ml/europlanet-ml-book), which serves as a tutorial and reference book for the activity in our work package.

Presentations and Workshops

Fireball-tracking networks around the world are assisting in the recovery of fragments of fresh meteorites and understanding where in the solar system they originated. In collaboration with NA2, the ML WP organised the second and third workshops in this series of four, which were held on 4-5 February 2022 (virtual) and on 13-14 August 2022 (hybrid). These workshops bring together observers from different fireball networks, along with ML experts, to discuss how ML can support the fireballs community and to advise on handling the data collected.

Four ML pipelines have been presented in three workshops during EPSC2022 - the pipeline for the IAP-CAS boundaries science case, the pipeline for the GMAP mounds science case, as well as two pipelines for the GMAP pits science case. All of the pipelines are available on GitHub.

Presentations with results of the science cases are mentioned in the section about the individual science cases.

Collaborations

We continued our collaboration with two research projects at the IWF.

Furthermore, we continued our collaboration with the EU Horizon 2020 project EXPLORE and we are further investigating the possibility of integrating our ML pipelines into the EXPLORE platform.

Task 3 - Data Pre-Processing, ETL and Feature Engineering

The aspects of data pre-processing and feature engineering are covered in the descriptions of the work for the individual science cases. Most science cases thereby utilize standard pre-processing methods or work on the raw data through end-to-end learning. However, we also explore new routes to automate pre-processing. For example, the GMAP Mounds science case utilizes data augmentation in the form of generative adversarial networks to overcome data sparsity. Details on the pre-processing conducted can be found below.

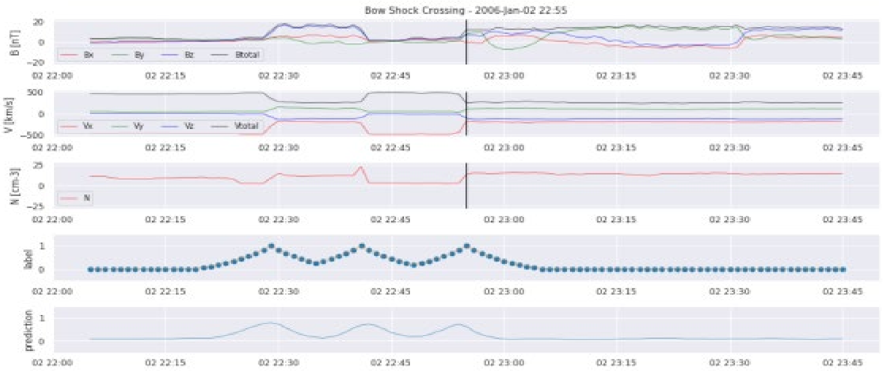
Task 4 - Time-based Signal Analysis and Automatic Classification

IWF ICME Science Case (Automatic Detection and Classification of Boundary Crossings in Spacecraft in situ Data)

Planetary magnetospheres create multiple sharp boundaries, such as the bow shock, where the solar wind plasma is decelerated and compressed, or the magnetopause, a transition between solar wind field and planetary field.

There have been several quite successful attempts to automatically segment in situ time series. Labeling the different regions such as the magnetosphere, the magnetosheath and the background solar wind, the segmented maps were subsequently used to detect boundary crossings and build a database accordingly. From an exploratory point of view, we were interested in whether it would be possible to train a convolutional neural network on a catalog of bow shock crossings to obtain these directly, without the need for a fully segmented time series.

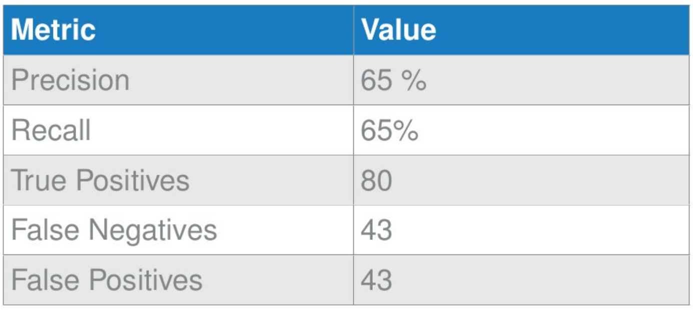
So far, we have started to develop a pipeline using only magnetic field and components, ion bulk-velocity and components, ion density, parallel ion temperature and perpendicular ion temperature from the Cluster 1 spacecraft, resampled to a 1-minute frequency. To account for the huge data imbalance, parts of the data, where no bow shock crossings are expected (for example when the spacecraft is in the night side part of the magnetopause or too far away in the solar wind) were removed. Since the temporal expansion of a bow shock crossing is quite limited, the labeling of the data had to be conducted thoughtfully. We decided on a parameter between 0 and 1, which simultaneously defines if a given time
frame contains a bow shock crossing and how far from the center it occurs.



The predicted label is clearly increasing for times when bow shock crossings occur. Thus, a peak detection algorithm can be used to extract a list of crossings. Even though precision and recall need to be improved, first results are promising and lead to the next steps:

- train on more data from different spacecraft
- use non-resampled datasets
- include additional features
- tune hyperparameters
- further experiment with model architecture
- cross-validation

Metrics for a random split of the data can be seen in the table below.



Metric	Value
Precision	65 %
Recall	65%
True Positives	80
False Negatives	43
False Positives	43

The ML pipeline is available on our GitHub repository.

This science case was presented at the EGU22 and the ESWW2022. Further, the results of this science case were published in the journal „Space Weather“:
Rüdisser, H. T., Windisch, A., Amerstorfer, U. V., Möstl, C., Amerstorfer, T., Bailey, R. L., & Reiss, M. A. (2022). **Automatic detection of interplanetary coronal mass ejections in solar wind in situ data.** *Space Weather*, 20, e2022SW003149. https://doi.org/10.1029/2022SW003149

LMSU Boundaries Science Case

We developed an efficient method to detect automatically the bow-shock and magnetopause boundary crossings using data from the MESSENGER magnetometer. To this end, we first prepared the data suited to Machine Learning. Next, we experimented with several ML models, specifically neural networks to find a usable baseline. Next we devised an Active Learning (AL) approach to select only the most informative orbits in the training set by using an uncertainty criterion. Using this strategy we were able to find a generalisable model with only 10% of the available data. A framework with these models was published and made available open-source for the community to experiment with on similar tasks. Results from this work were published in EGU, EPSC, ML-Helio 2022, and ECML-PKDD 2022.

We further extended this Active learning approach by augmenting it with a Drift Detection strategy, such that the data sampler would first detect a distributional shift in the data, and then use the entropy based criterion within the corresponding drift to order the most informative orbits. This further reduces the number of training orbits required, significantly outperforming random sampling. The results from this paper are in process of being assimilated into a paper to be submitted as invited contribution in an AGU journal.

The ML pipeline is available on our GitHub repository.

This science case was presented at ECML PKDD 2022 and published in the proceedings of this conference:
Julka, S., Kirschstein, N., Granitzer, M., Lavrukhin, A. & Amerstorfer, U. V. (2022). Deep Active Learning for Detection of Mercury's Bow Shock and Magnetopause Crossings. *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*. https://2022.ecmlpkdd.org/wp-content/uploads/2022/09/sub_1177.pdf

IAP wave emissions science case

This science case targets plasma wave identification from time-frequency spectrograms, specifically the electromagnetic whistler-mode "chorus" emission frequently observed in the inner magnetosphere of the Earth and other planets. The wave emissions typically occur as structured or unstructured features with visible boundaries in the time-frequency domain. We have created and delivered a training dataset of time-frequency spectrograms, 10 seconds each, generated from the data of the Wideband receiver on board four Cluster spacecraft. There are more than 4000 events irregularly observed while spacecraft crossed the terrestrial magnetosphere and the nearby solar wind. We have visually checked the data and classified the intervals based on whether the chorus emission is present (82% of events) or not (18% of events). The value of local electron cyclotron frequency is included in the data for easier identification. The dataset is provided in the form of Python data structures which can be used for both supervised and unsupervised machine learning.

Uni Passau received the data from the partners in Prague. Uni Passau preprocessed the data, and conducted the basic explorative analysis. The task is to segment the pixels with the whistler waves in the spectrograms, and only have image level labels available; as a first task in feasibility, Uni Passau built a classifier that classifies the image sample into a binary category of interest. This classifier provides an accuracy of about 70 %. As the segmentation task itself needs to be unsupervised, ie. without labels, Uni Passau is investigating an approach that might learn a similarity metric from the data and disentangle the representation space into relevant and irrelevant parts. This will be the major focus of this year's development.

Task 5 - Images and Other (General) Classification Tools

GMAP Mounds Science Case

Before the commencement of the previous year, we had already developed a basic segmentation pipeline using generative models (Image to Mask Autoencoders) to perform image segmentation and detect mound like features in the DTMs. As the data provided was severely limited we initially attempted to generate simulated samples as an attempt to augment the training set. But that did not improve the segmentation. Finally, we augmented the training set with engineered features viz slope, aspect and hillshade and achieved a reasonable performance in the segmentation of mounds. Owing to the lack of a proper validation set, it cannot be confirmed if the model will generalise well to unseen images.

In a parallel line of experiments, we investigated the representation space learnt by the generative models, wherein we devised a method to disentangle the mound-specific representation from the non-mound representation, in order to perform controlled simulation. This would be useful to improve searchability in compressed representations, but it is yet to be determined if this approach can help directly in the main goal of segmentation.

Results from this work were presented in EPSC, EGU 2021, and the segmentation pipeline was demonstrated in a workshop in EPSC 2022.

The ML pipeline is available on our GitHub repository. The ML pipeline was presented during a workshop at the EPSC2022.

GMAP Pits Science Case

We are planning the next developing steps for the DeepLandforms tool; for instance, we want to generalize it further by providing further baseline configuration. For instance, a configuration for PyTorch and another for Tensorflow Python packages.
We are preparing a newer dataset to be tested with the updated tool.
DeepLandforms has been presented with a live demo in a splinter-session at the Europlanet Science Congress 2022, held in Granada.
The paper presenting DeepLandforms, was accepted on December 2022, and is available on https://doi.org/10.1029/2022EA002278.
The code is completely available on the EPN-ML GitHub.
Constructor University (formerly Jacobs University) press release: https://www.jacobs-university.de/news/researchers-develop-ai-method-mapping-planets

DLR Surface Composition Science Case

During the last year, DLR organized the code repository to be self-sufficient from data source to end result. The DLR ML team produced a complete report (PDF/HTML) present in the repository and published with Elsevier. The aim is to make the user able to reproduce the complete work just using the information contained in the repository. A video tutorial on DLR use case for the GMAP Winter School has been released. The presentation is in the DLR repository as well, and the video is available on GMAP server (https://www.planetarymapping.eu/).

INAF spectral analysis for planetary minerals case

We analyzed the best spectra data to provide to the project and implemented a procedure to format the data in a standard way using JSON format for data storage and transport; in this way the communication of the methods of reading and managing the dataset will be simple, the metadata necessary for their understanding will also be stored in the dataset. We test some ML algorithms on the selected dataset, to test if the dataset can be used for an ML analysis.

AOP Asteroid/meteor Science Case

During this reporting period, it was determined that Machine Learning would not be of benefit to the science cases originally proposed by AOP. Following some conferring within WP10, a re-defined science case was formulated aimed at the classification of meteor lightcurves. AC is presently in collaboration with Andreas Windisch (FH JOANNEUM) to take this forward to implementation. Work by AC in the last few months of 2022 focused on extracting the lightcurves from the raw data and already several thousand lightcurves have been made available to Dr Windisch and his group for pre-processing.

Task 6 - Virtual Access and Interfaces

The Machine Learning Portal provides the public point of entry to our ML activities. We continuously update the content on the portal. We also improved the ML Portal structure according to the comments of the VA review board.

In the past year, we have been working on finalising and disseminating the APP (Analysing Planetary Pits) tool. APP is a Python framework for automatically deriving apparent depth profiles of Solar System pits by measuring the width of their shadows. It uses image segmentation to separate cropped satellite images (single- or multi-band) into shadow pixels or non-shadow background and calculates a profile of the apparent depth (h) of a pit – the depth at the edge of the shadow – along the entire length of the shadow. The testing of the shadow extraction is complete, proving that k-means clustering with silhouette analysis was the most accurate method. APP has been presented at a number of conferences and forums over the past year (EAS Annual Meeting, RAS National Astronomy Meeting, Congress on Geomorphology), including hands-on sessions at the Europlanet Science Congress

2022 in Granada where participants got to use APP in practise for the first time. A paper has been written describing APP, which has now been submitted to the journal - Royal Astronomical Society's Techniques and Instruments (RASTI). The short-term plan is to make the tool publicly available in the ACRI-ST GitLab, the Europlanet GitHub and the EXPLORE platform.

c. Impact to date

We have a rising number of visitors on our ML Portal. At different occasions, e.g. conferences, we have presented results of our science cases as well as our ML activities in EPN-2024-RI.

We have published three publications with ML contributions and results of our science cases.

We have organized and convened four conference sessions specifically dedicated to ML in planetary sciences and heliophysics (and we will organize such sessions again in 2023).

Five workshops were conducted in the course of EPSC2021 and EPSC2022 to introduce our ML pipelines to the scientific community.

We have a growing number of ML pipelines on our public GitHub repository.

d. Summary of plans for Year 4

Currently, we are drafting publications with the results of at least two science cases (GMAP pits, IAP-CAS boundaries). We will finalize the integration of first data sets of our science cases into VESPA. Further, we will integrate the first ML pipelines into SPIDER and the EXPLORE platform.

There will be the fourth and last Fireballs workshop, co-organized by ML. Furthermore, we will again have the session „Machine Learning in Planetary Sciences and Heliophysics" at EGU 2023. For this session, 24 abstracts were submitted.

Finally, we will finalize our work on the remaining science cases and publish the final version of our Jupyter book, containing documentation and tutorials about our work.

2. Update of data management plan

The data management plan will be updated in order to incorporate the comments raised by the VA review board.

3. Follow-up of recommendations & comments from previous review(s)

We have answered the issues raised in the VA review board report in a collected answer of all VAs.