D1.3 Documentation of individual datasets (CML, PWS, SML) shared in the repository

Version 1.0

Working Group	WG1
Deliverable	D1.3
Date	28/11/2023
Version	1.0
Authors	Roberto Nebuloni (CNR-Italy) and Martin Fencl (CTU-Czech Rep.)
Reviewer	Vojtěch Bareš (CTU-Czech Rep.)

Description	This document describes datasets of CMLs and SMLs shared by OpenSense members within the OpenSense research community. Some of them will be used to test the algorithms for processing opportunistic sensing data and assess data applications.
Keywords	Repository, datasets, OpenSense, CML, SML, PWS, HD Rain, Meteo France

About OpenSense (COST Action CA20136). *OpenSense brings together scientists investigating different opportunistic sensors (e.g. microwave links, citizen science), experts from weather services, and end-users of rainfall products to build a worldwide reference opportunistic sensing community. The overarching goals of the COST are to overcome key barriers preventing data exchange and acceptance as hydrometeorological observations, define standards to allow for large-scale benchmarking of opportunistic sensing precipitation products and develop new methods for precipitation retrieval, coordinate integration of the opportunistic observations into traditional monitoring networks, and identify potential new sources of precipitation observations. Further details can be found here:*

Funded by the European Union

www.cost.eu
www.opensenseaction.eu

Table of contents

Glossary

CML	Commercial Microwave Links
csv	Comma-separated values (filename extension associated with text files)
DM.N	Deliverable N of working group M
GPK	Grant period K
MoU	Memorandum of Understanding
NDA	Non-Disclosure Agreement
nc	Filename extension of the NetCDF (Network Common Data Form) data format
OpenSense	Opportunistic precipitation sensing network
OS	Opportunistic Sensors
PWS	Personal Weather Stations
SML	Satellite Microwave Links
WG	Working Group

1. Introduction

The official Opensense Deliverable D1.3, entitled "**Documentation of past datasets shared in the repository"** is an output of the activities of curation, sharing and standardization of individual datasets collected by OpenSense members in the frame of their past or current research work. The above tasks are being carried out by OpenSense Working Group 1 (WG1). Details about WG1 activities, milestones and deliverables due are reported in the OpenSense MoU [1] and are listed in Table 1.

D1.3 builds from the internal deliverable D1.1 entitled "Repository for individual OS datasets" [2], which was delivered at the end of GP1. It was the first output of the activity of dataset curation and sharing. D1.1 describes the GDrive repository where the shared datasets are stored, lists the available datasets and provides some basic information for each one, such as type of sensor (i.e. CML, SML, PWS, etc), data owners and format. Sections 2 and 3 of this deliverable are taken from D1.1. Table 2 in Section 3 summarizes all the datasets currently shared on Gdrive adding information about the size of data (number of sensors and duration of observations). Some of the above databases are more than samples, as they cover relatively large areas and come together with conventional data. Section 4 details the structure of the most comprehensive OS dataset so far available on GDrive, that is, the SML dataset provided by the French company HD Rain. Despite not being stored on GDrive, concurrent conventional data provided by Meteo France are available as well and can be shared within the OpenSense community.

Table 1 Timetable of WG1 activities, milestones and deliverables according to OpenSense MoU.

	year 1				year 2				year 3				year 4			
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
WG1 - data management and standardization																
Defining common data and metadata format for data exchange				D1.1												
Sharing and curration of individual datasets				D1.2		D1.3										
Gathering large-scale OS dataset												D1.4				
Complementing of OS datasets with standard observations																
Compiling joint benchmark datasets								D1.5								
Establishing + maintaining operational access to OS data																
Project Milestones				M1		M2		M3		M4			M5			
WG deliverables																

WG1 milestones
M2 large-scale dataset in standardized format available for benchmarking of algorithms
M5 - transboundary OS precipitation product available
M6 - operational access to subset of OS data established

WG1 deliverables
D1.1 Repository for individual OS datasets
D1.2 White paper on data standards/formats for investigated types of OS sensors
D1.3 Documentation of past datasets shared in the repository
D1.4 Large-scale OS dataset completed by standard observations + report
D1.5 Benchmark dataset and documentation in the form of a report
D1.6 Documentation for accessing operational OS data

2. Data sharing policy

The ultimate goal of the OpenSense WG1 activity is to unlock large OS datasets and make them available to everyone in an open data repository. To this aim, OpenSense members were asked to share a sample of the data they are using or used in their past research activity on the topic of OS.

However, for some types of OS data, there are issues with data ownership. For instance, CML data are generated by terrestrial communication networks for purposes of link quality control. A similar scenario regards SML signals generated for telecommunication services. The owners of

these data are the mobile or satellite operators. In other cases, for instance when SML data are broadcasted by GEO satellites to provide TV-SAT services, they signals received at the ground are available to anybody. This is also the case of GNSS signals. Despite CML and SML data generally do not contain information about user traffic, hence they do not involve privacy issues, the associated metadata (e.g. position of user terminals or base stations), which are necessary to process the data, are often classified as confidential by the owners. Therefore, these data are usually shared by the owners (sometimes not for free) with researchers for purposes of scientific project activities, following the signature of an NDA. On the other hand, other types of OS data such as the ones collected by PWS are generated by non-professional grade weather equipment owned by private citizens and made available on public platforms where people can download them for free (see for instance Wundermap by Weather Underground [3] or Netatmo platform [4]) or paying some fee.

Given the above restrictions on open OS data sharing, it was agreed to create, as an intermediate step, a repository accessible only to the OpenSense community (more than 100 researchers at the moment). The repository is on Google Gdrive and it currently contains dataset samples provided by several research groups, as detailed in the next section. These sample datasets include CML, SML and PWS data with different formats and a different organization of the data. They provided some useful inputs to the WG1 activity of standardization of data formats, which has been finalized and summarized in the Opensense Deliverable D1.2 entitled “White paper on data standards/formats for investigated types of OS sensors” [5].

The efforts of WG1 towards publishing more datasets under licenses enabling their sharing and reuse will continue to be coordinated with WG4. In the meantime, we have created an OpenSense community collection on the Zenodo repository which associates already published OS datasets with OpenSense Action [6].

3. Structure of the repository accessible to OpenSense members

The repository in the WG1 shared folder (`OpenSenseAction > WG1 > Data_Repository`). At the date when this document was issued, the repository included 12 subfolders each storing sample OS data.

A readme file in the folder `Data_Repository` provides the guidelines to the OpenSense members for uploading their sample data on Gdrive. Specifically:

- Sample data folders have straightforward names in the following format: `Mysensor_mycountry_myinstitution`. For instance, one of the folders is named `CMLs_Sweden_SMHI` and stores CML data collected in Sweden by the Swedish Meteorological and Hydrological Institute.
- The data stored in each subfolder include the following elements:
    - Data files, i.e. the set of raw data, typically in the form of time series (e.g. power levels and the corresponding time stamps in the case of CMLs) that can be processed to extract rainfall information, i.e. intensity or accumulated precipitation in a given lapse of time.
    - Metadata files if any, that is, all the data necessary to extract rainfall information from the raw data (e.g. link position and frequency in the case of CMLs).
    - A short readme file with instructions on how to read data and metadata files as well as with a short statement about the terms of use of the data.

Table 2 lists the available datasets. The dataset #6 is open and can be used by everybody and processed by a Python tool available on the github platform (see Table 2). OpenSense members of the Swedish Meteorological and Hydrological Institute have recently published a larger open dataset including CMLs alongside rain gauge and radar data [7].

Table 2 List of OS data samples stored in the official Google GDrive repository of OpenSense action (as of 27 November 2023) and available to Opensense members.

	Folder	Owner	Format	Sample sze and extra info
1	CMLs_china_HHU	Hohai University	csv	1 link in Nanchang (China), 1 day. No rain.
2	CMLs_Czechia_CTU	Czech Technical Univ.	csv netcdf	2 links in Prague (Czech Rep.), 3 days. Rain.
3	CMLs_Germany_KIT	Karlsruhe Institute of Technology	csv	5 links, 2 days. Rain. The data are a subset of the CML dataset published within pycomlink under a BDS-3 License. For more information on the license visit https://github.com/pycomlink/pycomlink/blob/master/LICENSE
4	CMLs_Israel_TAU	Tel-Aviv University	csv	1 link, non consecutive days from 1-Jan-2013 to 1-Feb-2013. Rain. Data format is in Min-Max.
5	CMLs_Italy_CNR	National Research Council of Italy	csv	3 links in Valmalenco (Northern Italy), 1 day. Rain. Data format is in Min-Max.
6	CMLs_Netherlands_KNMI_WUR_TUDelft	Royal Netherlands Meteorological Institute; Wageningen University & Research; Technical Univ. of Delft	csv	2500 links covering The Netherlands, 21 days. Rain. Data format is in Min-Max. The dataset can be used with the open-source R package RAINLINK available at https://github.com/overeem11/RAINLINK. These data are open and can be publicly used.

7	CMLs_Sweden_SMHI	Swedish Meteorological and Hydrological Institute	nc csv (metadata)	2 links, 1 day. Rain. The data are a sample from the openly available database OPENMRG accessible at https://doi.org/10.5281/zenodo.7107689
8	PWS_SBLA_Cyprus		csv	10 stations in LImassol (Cyprus), 1 month Data are 1-min precipitation.
9	PWSs_Germany_UniStuttgart	Univ. of Stuttgart	HDF5	Only a template of their PWS data
10	SMLs_dataset_by_SRS_ARTYS	Artys s.r.l.	csv	26 links in Valpolcevera catchment area (Liguria region, Northern Italy), 14 days (6 rain events)
11	SMLs_Italy_UNIPI	Univ. of Pisa	csv	2 links in the cities of Pisa and Massa (Tuscany, Central Italy), 5 days (3 rain events).
12	SMLs_France_HDRain	HDrain	csv	215 links in Southern France, 5 months (Aug 2022-Dec. 2022) and 15 links in Southern France, 4 days in 2021 Rain.

4. HD Rain SML Dataset

The dataset has been provided by Dr. François Mercier-Tigrine of HD Rain, a French company that commercializes weather-products [8]. The dataset includes five months of continuous data from August to December 2022 collected by an overall 215 SML sensors deployed in Southern France (Fig. 1). The sensors receive the signals transmitted by the GEOstationary satellites Astra19 and HotBird13E at frequencies ranging in the upper part of the Ku band, i.e. between 11.7 and 12.75 GHz, with horizontal polarization.

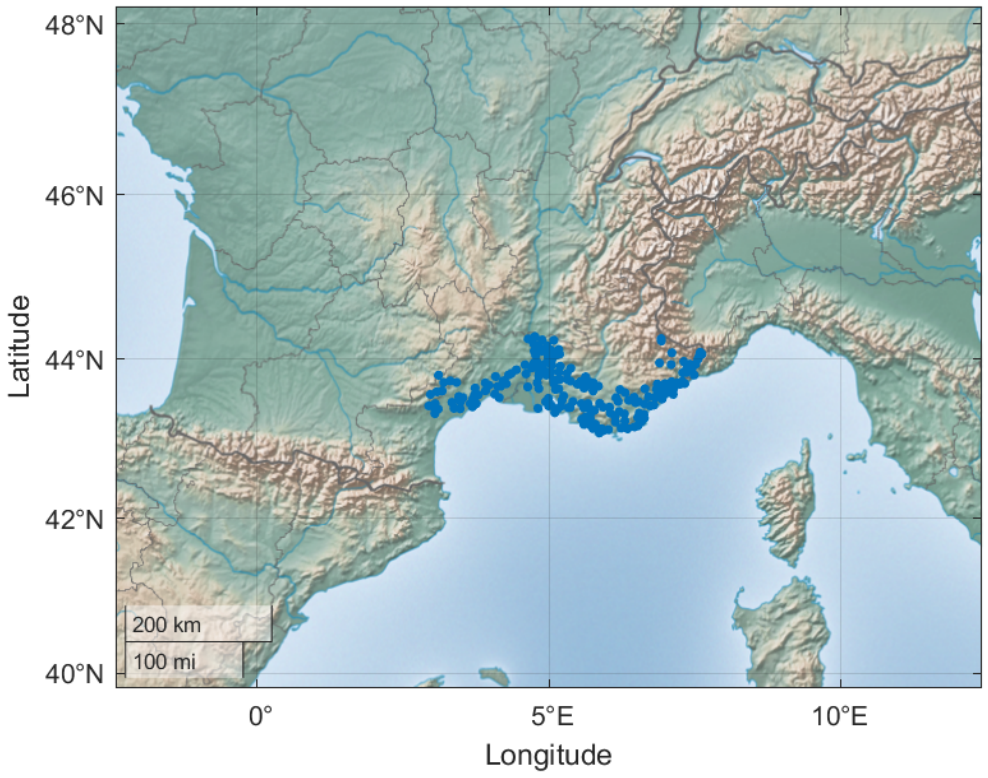


Fig. 1 Location of the 215 HD Rain ground receivers in Southern France

The HD Rain dataset is stored in the WG1 shared folder

`OpenSenseAction > WG1 > Data_Repository > SMLs_France_HDRain > 2022`

where there are two files

- a csv file named

  `metedata_2022.csv`

  that stores metadata, i.e. information about sensor location and sensor features.
- a zip file named

  `data_hdrain_2022.zip`

  that stores 215 files .csv files, one for each SML sensor, with the raw data, i.e. the values of the received signal level (RSL) at ground.

Each file has a conventional name in the form

`FRA-NNNNN.csv`

where NNNNN is a five-digit integer. The sensor identifier is the entire filename (except the extension) and it is the first record of each line in the metadata file.

Files in the csv format are basically text files, hence they can be opened with a text editor as Microsoft Notepad, or with a spreadsheet as Microsoft Excel, or, again, with tools for scientific data computing as Matlab. For instance, an HD Rain data file can be loaded into MATLAB prompting the following simple instructions:

```
file_name='FRA-00038.csv';
fid=fopen(file_name);
C = textscan(fid, '%s%f','HeaderLines',1,'Delimiter',';');
fclose(fid)
```

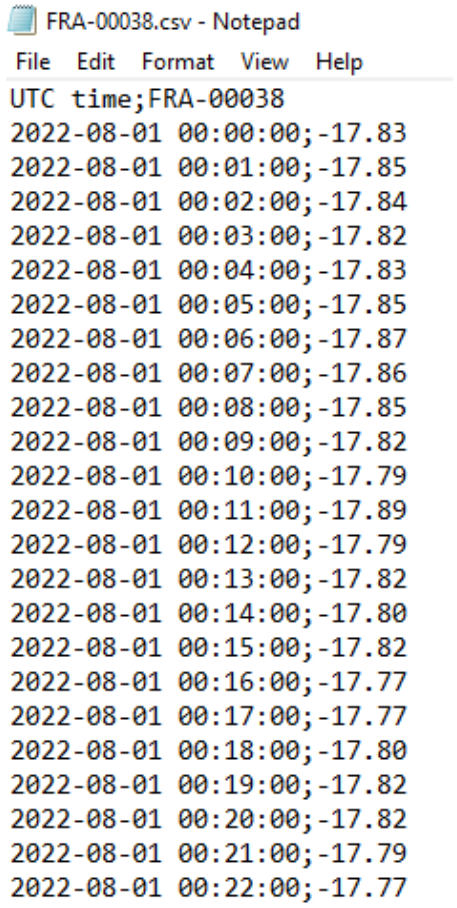
where `C` is a cell-array of two elements corresponding to the time stamp record and the RSL record, respectively.

Fig. 2 shows a screenshot of how the csv files with RSL data look like, when they are visualized by a text editor (a) or by the Microsoft Excel spreadsheet. There are two columns separated by commas, i.e.:

- time stamp in the format yyyy-mm-dd HH:MM:SS in UTC
- Received Signal Level (RSL) in dBm.

The first line of the file is a header.

(a) Notepad window



(b) Excel window

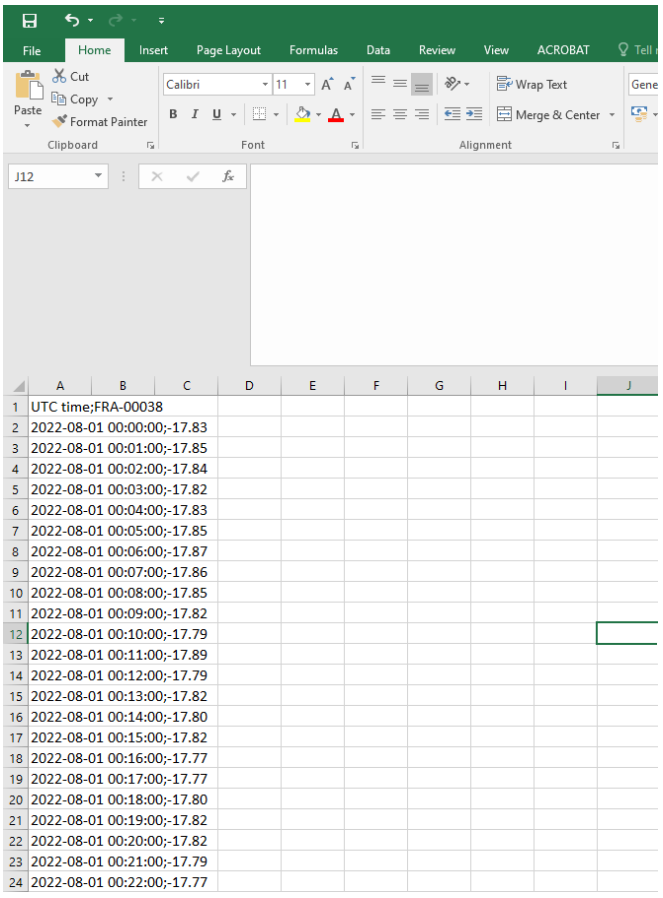


Fig. 2 Example of a csv file storing HD Rain SML data

The content of the HD Rain metadata file is shown in Fig. 3 After the header line, there are 215 lines, one for each sensor. Each line stores with the following records:

- sensor identifier (i.e. the name of the corresponding RSL data file)
- latitude of the sensor (deg)
- longitude of the sensor (deg)
- elevation a.m.s.l. of the sensor (m)
- satellite azimuth (deg)
- satellite elevation angle (deg)
- satellite name

metedata_2022.csv - Notepad

File Edit Format View Help

```
station ID;latitude;longitude;elevation;satellite_azimut;satellite_elevation;satellite
FRA-00002;44.253;4.646;55;159.862;36.950;Astra 19
FRA-00008;43.353;6.206;262;161.696;38.300;Astra 19
FRA-00018;43.689;7.046;300;162.959;38.154;Astra 19
FRA-00021;43.567;7.131;8;171.517;39.409;HotBird 13E
FRA-00022;43.580;7.123;16;171.507;39.394;HotBird 13E
FRA-00025;43.633;7.095;64;171.475;39.333;HotBird 13E
FRA-00028;43.616;7.072;134;162.974;38.237;Astra 19
FRA-00029;43.885;4.860;34;160.029;37.389;Astra 19
FRA-00030;43.895;4.708;25;159.825;37.337;Astra 19
FRA-00034;43.793;4.828;42;168.277;38.840;HotBird 13E
FRA-00035;43.807;4.956;56;168.462;38.846;HotBird 13E
FRA-00036;43.710;4.802;11;159.891;37.554;Astra 19
FRA-00037;43.867;4.911;45;160.092;37.422;Astra 19
FRA-00038;43.839;4.837;19;168.300;38.791;HotBird 13E
FRA-00039;43.745;5.080;98;168.625;38.932;HotBird 13E
FRA-00040;43.699;5.086;111;168.624;38.983;HotBird 13E
FRA-00041;43.636;5.171;100;168.731;39.065;HotBird 13E
FRA-00042;43.629;5.114;74;160.288;37.723;Astra 19
FRA-00045;43.393;4.814;1;159.800;37.885;Astra 19
FRA-00047;43.659;4.639;5;167.982;38.954;HotBird 13E
FRA-00048;43.639;4.830;25;168.248;39.008;HotBird 13E
FRA-00049;43.424;5.049;23;160.130;37.918;Astra 19
FRA-00050;43.518;4.992;0;168.453;39.164;HotBird 13E
FRA-00051;43.476;4.948;15;160.010;37.836;Astra 19
```

Fig. 3 The csv file storing HD Rain SML metadata.

The transmitting satellites are in GEOstationary orbit. Specifically, the satellites used by HDrain sensors are Astra19 and HotBird13E, which host payloads for TV-sat broadcasting. A list of active satellites as well as of their main features is in [9]. Each sensor receives the signal from either Astra19 or HotBird13E. The RSL values are available every 1-min. They represent the average received signal over 1-min. An example of RSL time series is shown in Fig. 4

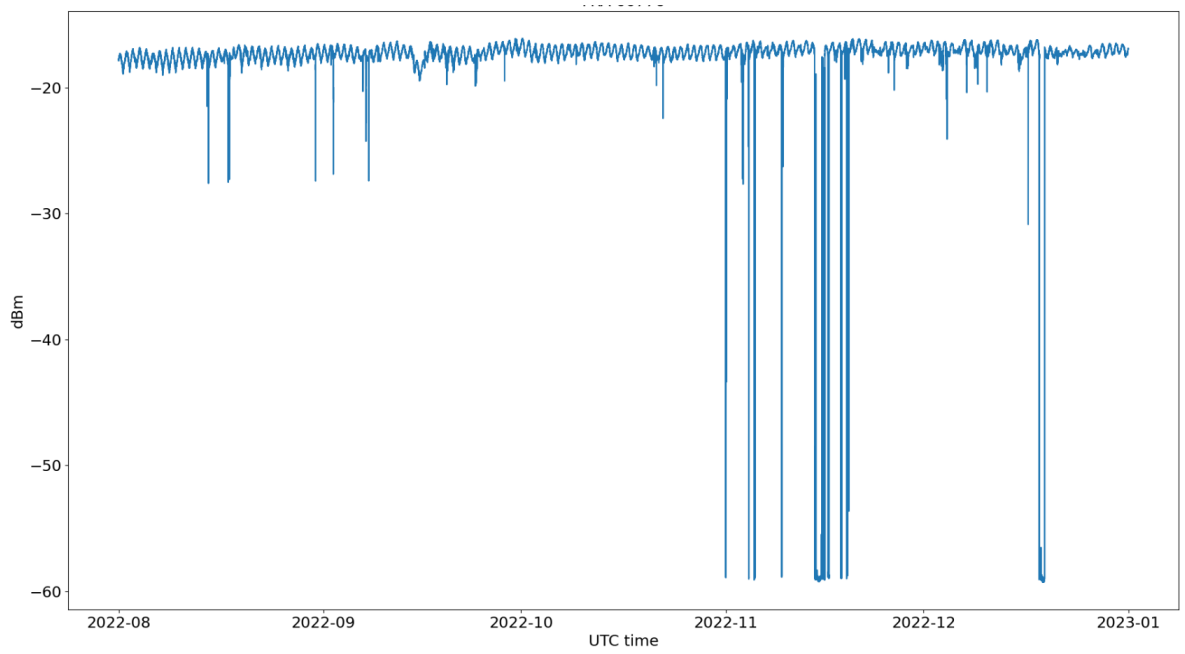


Fig. 4 Time series of raw RSL data for the sensor FRA-00038 over the available five-month period of time.

The HD Rain database is complemented by two sets of conventional meteorological data, courteously provided by Meteo France, under a research license agreement valid until the end of the OpenSense action (31 October 2025). Please note that conventional data are not shared on GDrive even though they are available to the OpenSense members upon request to WG1 leaders.

The conventional dataset includes

- Radar data
- Ground data (weather stations)

The conventional data can be used for calibration and/or validation of the SML data. Both radar and ground station data include the five months from August to December 2022.

The radar data are provided in Netcdf4 format (.nc file extension). Two different products are available:

- "COMEPHORE" Mosaic of accumulated precipitation during 1 hour in 1 km resolution over France.
- French Mosaic of accumulated precipitation during 5 min in 1 km resolution over France.

In case of COMEPHORE data, maps are stored in five large monthly files with the following names:

```
comephore_1km-1h_202208.nc
comephore_1km-1h_202209.nc
comephore_1km-1h_202210.nc
comephore_1km-1h_202211.nc
comephore_1km-1h_202212.nc
```

In the case of 5-min data, every map is stored in a separate file. The filename format is:

```
cumul_france_1536-1km-5min_yyyymmddHHMM.nc
```

where yyyymmddHHMM is a string with date and time (In UTC coordinates)

Radar data are gridded in a regular 1536 x 1536 matrix covering the European part of France (called 'Metropolitan France' by Meteo France). The spatial coordinates are polar stereographic coordinates. More details about Meteo France radar data can be found on the web [10]. A few examples of 1-h accumulated precipitation is in Fig. 5, while 5-min maps are in Fig. 6.

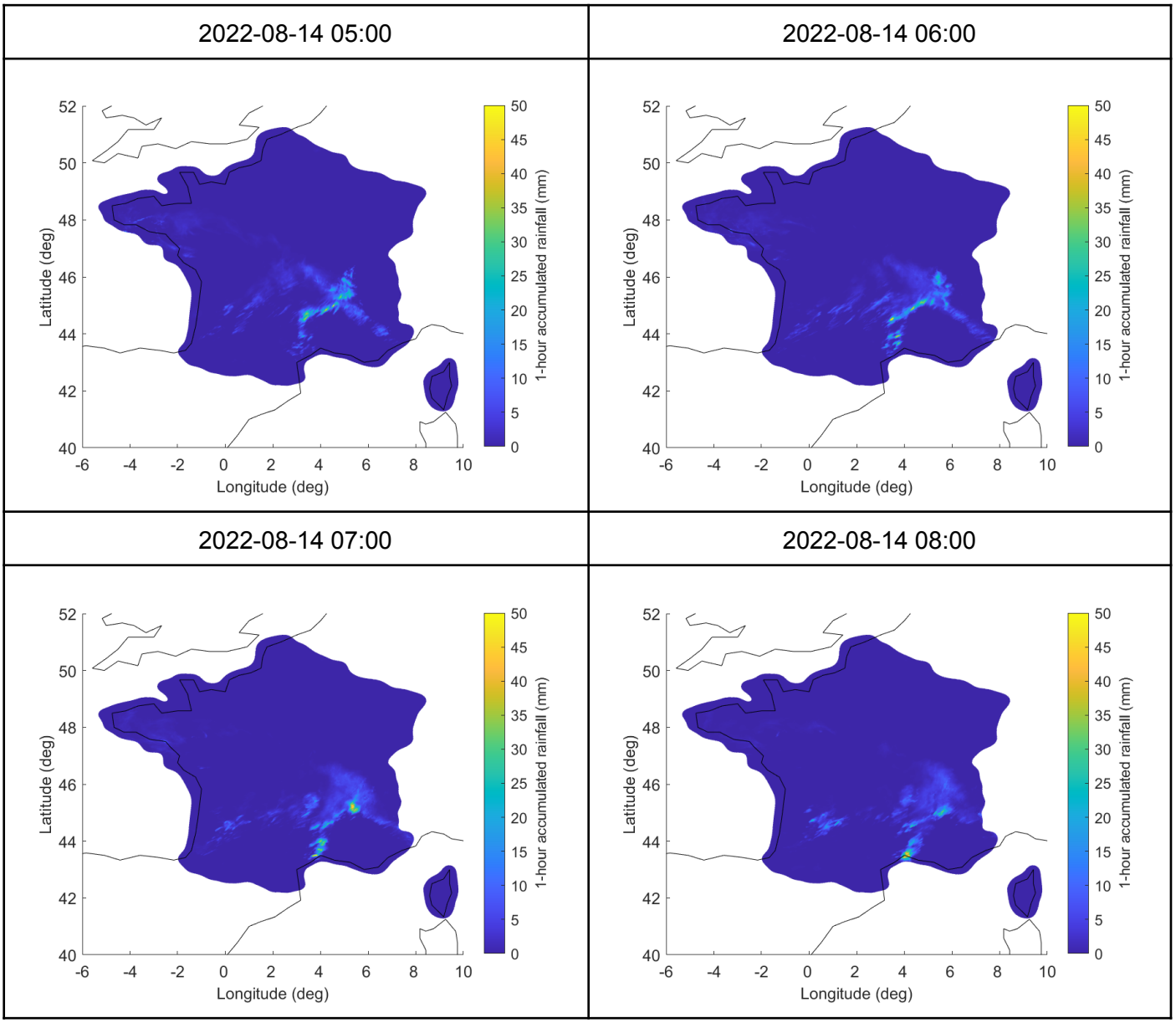


Fig. 5 A few examples of Meteo France Comephore radar data.

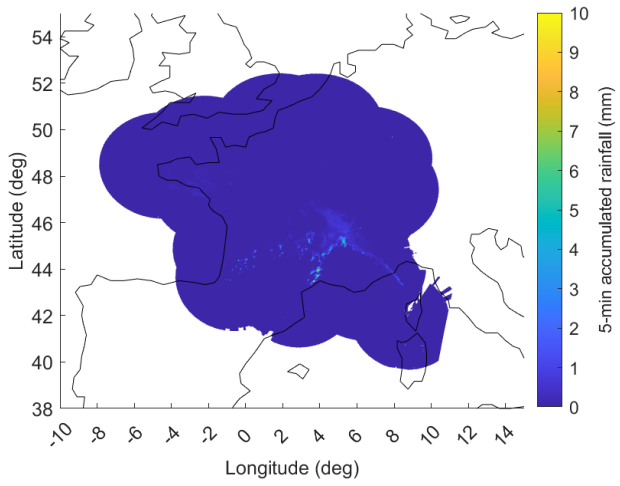
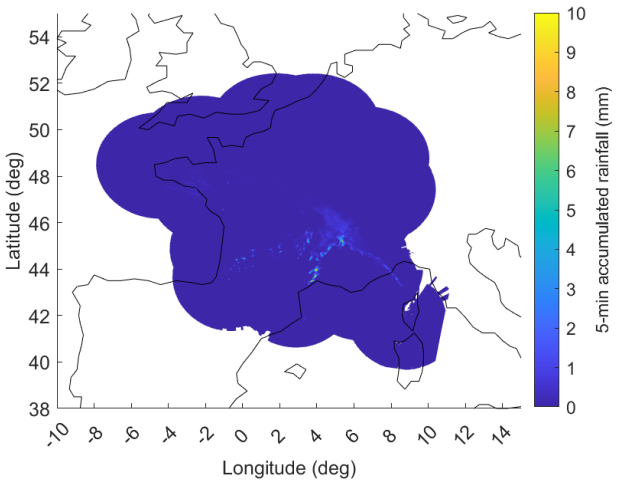
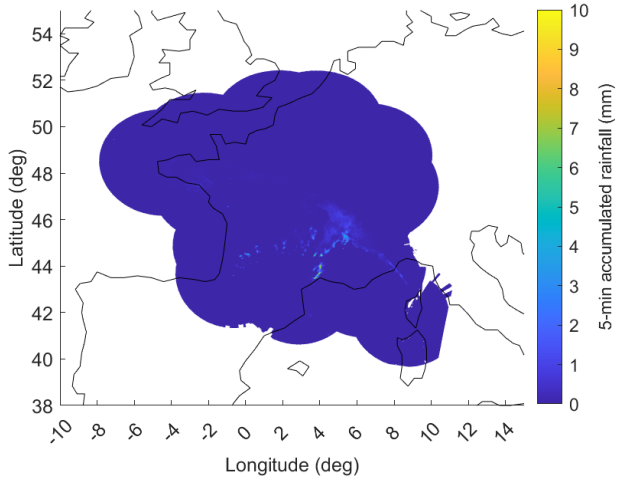
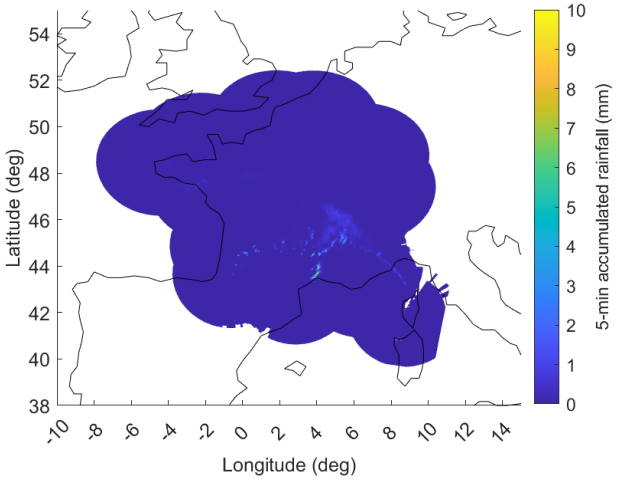
2022-08-14 07:05	2022-08-14 07>15
	
2022-08-14 07:25	**2022-08-14 08:35**
	

Fig. 6 A few examples of Meteo France 5-min radar data.

Data from 12 Meteo France ground stations located in the Departments of Southern France where HD Rain SMLs are installed are shown in Fig. The data collected by ground stations includes:

- atmospheric data every 1-hour, i.e. temperature, pressure, relative humidity, wind velocity and direction and 1-hour accumulated precipitation in the above 12 stations
- 6-min accumulated precipitation in 12 stations

Examples of 1-h data are reported in the following Fig. 8 and Fig. 9, while an example of 6-min accumulated precipitation is shown in Fig. 10.

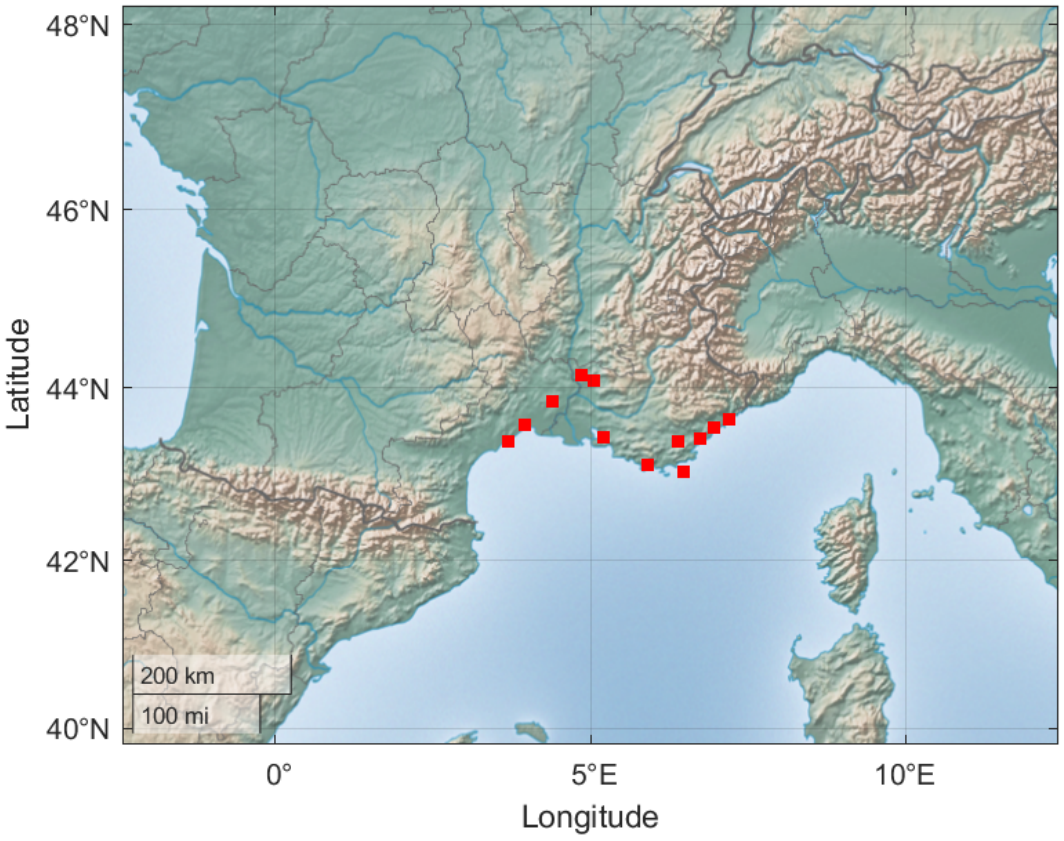


Fig. 7 Location of 12 available Meteo France ground stations in Southern France.

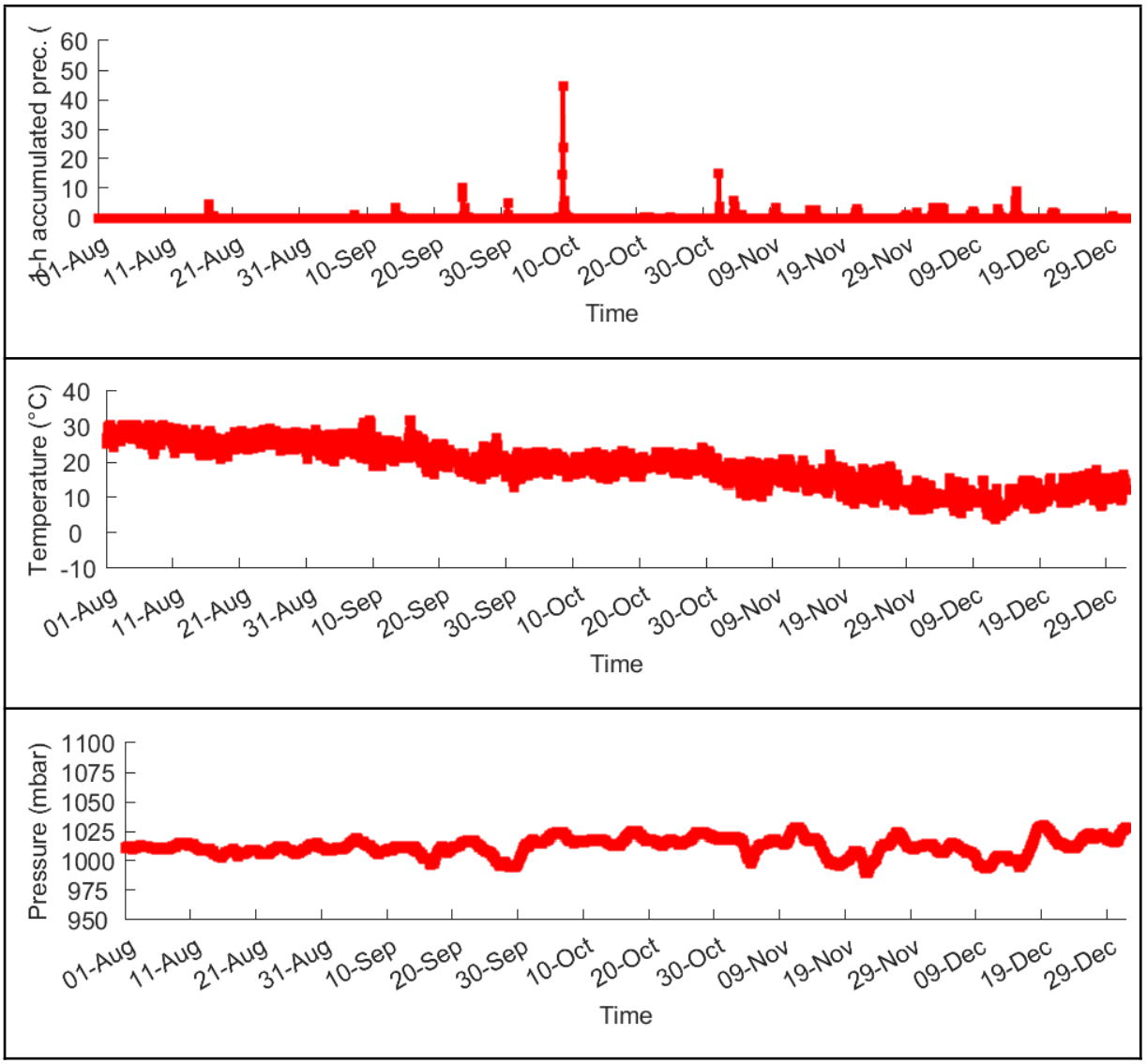


Fig. 8 Hourly ground station data measured by the Meteo France weather station in Nice in the period Aug-Dec 2022: 1-h accumulated precipitation, instantaneous temperature and pressure at the hourly timestamp.

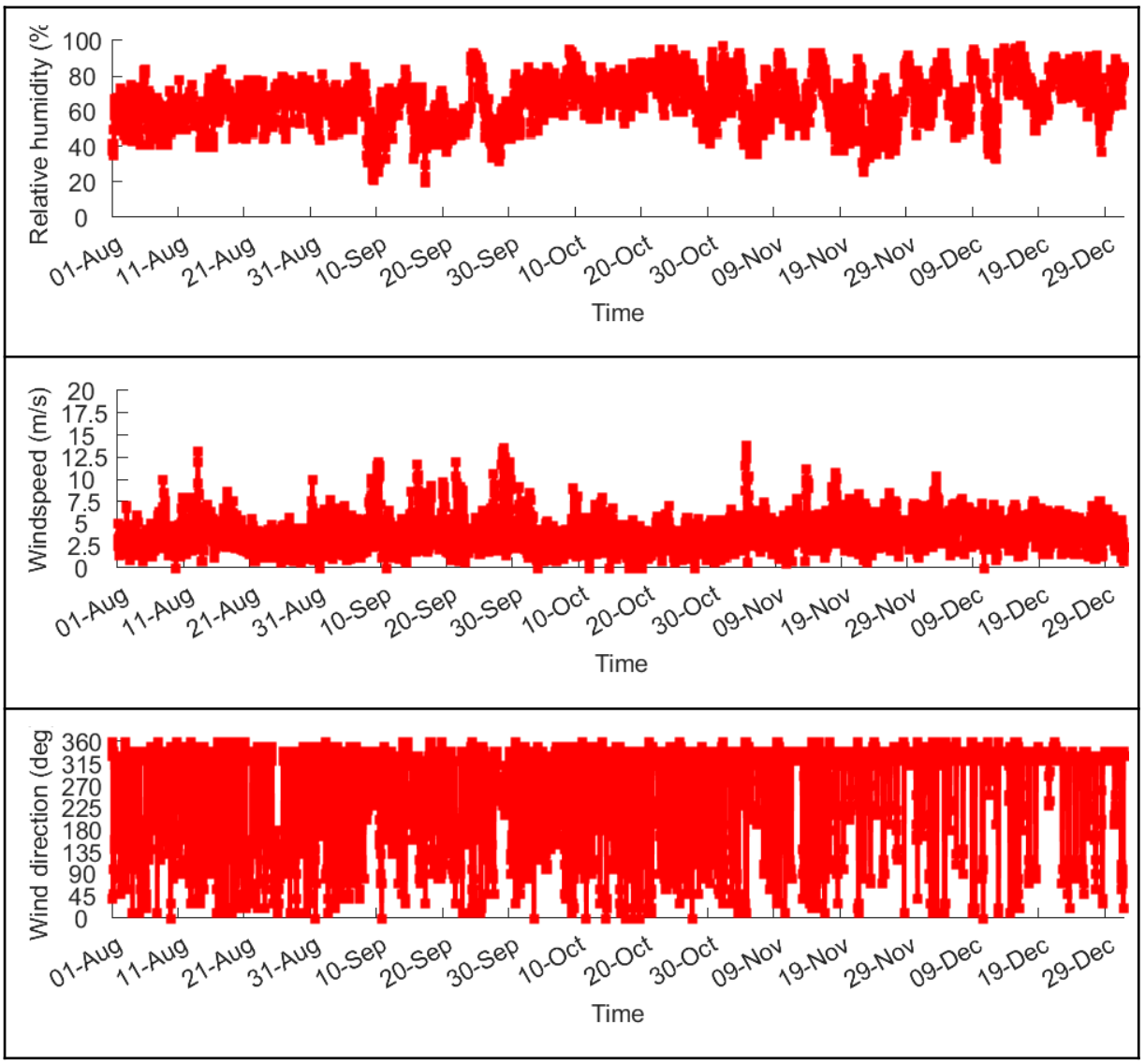


Fig. 9 Hourly ground station data measured by the Meteo France weather station in Nice in the period Aug-Dec 2022: relative humidity at the hourly timestamp, horizontal wind velocity at 10 m above ground averaged over 10-min before the time stamp, wind direction at 10 m above ground averaged over 10-min before the hourly time stamp.

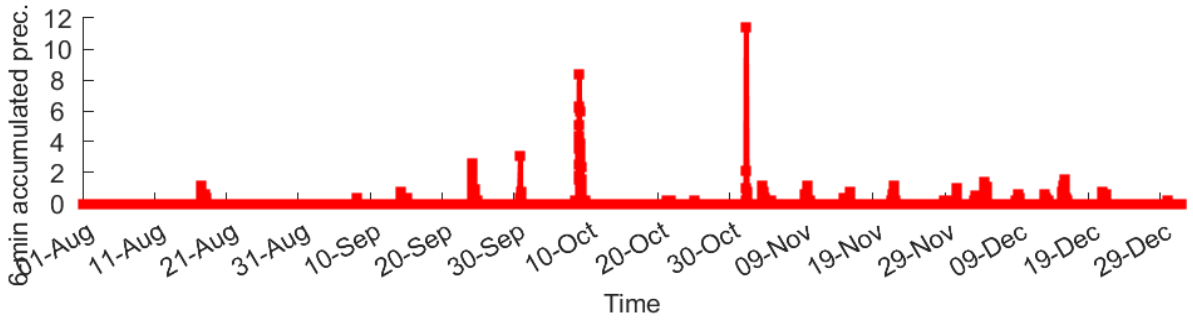


Fig. 10 6-min accumulated precipitation by Meteo France rain gauge in Nice in the period Aug-Dec 2022.

References

[1] Memorandum of Understanding for the implementation of the COST Action "Opportunistic precipitation sensing network" (OpenSense) CA20136, downloadable at https://www.cost.eu/actions/CA20136/

[2] "Repository for individual OS datasets", internal deliverable (D1.1) of the COST Action "Opportunistic precipitation sensing network" (OpenSense) CA20136,

[3] Weather Underground weather information platform, https://www.wunderground.com/wundermap

[4] Weathermap managed by the Netatmo company, https://weathermap.netatmo.com

[5] "White paper on data standards/formats for investigated types of OS sensors", official deliverable (D1.2) of the COST Action "Opportunistic precipitation sensing network" (OpenSense) CA20136,

[6] OpenSense community collection on Zenodo, https://zenodo.org/communities/opensense/about/

[7] Andersson, J. C. M., Olsson, J., van de Beek, R. (C. Z.)., and Hansryd, J.: OpenMRG: Open data from Microwave links, Radar, and Gauges for rainfall quantification in Gothenburg, Sweden, Earth Syst. Sci. Data, 14, 5411–5426, https://doi.org/10.5194/essd-14-5411-2022, 2022.

[8] HD Rain company official site, https://www.hd-rain.com/

[9] https://www.satbeams.com/channels

[10] AERIS home page, https://radarsmf.aeris-data.fr/en/home-page/