POLITECNICO DI TORINO
Repository ISTITUZIONALE

The Italian open data meteorological portal: MISTRAL

Original
The Italian open data meteorological portal: MISTRAL / Bottazzi, M.; Scipione, G.; Marras, G. F.; Trotta, G.; D'Antonio, M.; Chiavarini, B.; Caroli, C.; Montanari, M.; Bassini, S.; Gascon, E.; Hewson, T.; Montani, A.; Cesari, D.; Minguzzi, E.; Paccagnella, T.; Pelosini, R.; Bertolotto, P.; Monaco, L.; Forconi, M.; Giovannini, L.; Cacciamani, C.; Passeri, L. D.; Pieralice, A.. - In: METEOROLOGICAL APPLICATIONS. - ISSN 1350-4827. - ELETTRONICO. - 28:4(2021). [10.1002/met.2004]

Availability:
This version is available at: 11583/2939577 since: 2021-11-23T12:05:16Z

Publisher:
John Wiley and Sons Ltd

Published
DOI:10.1002/met.2004

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

12 October 2024

Received: 25 March 2021 | Revised: 18 May 2021 | Accepted: 19 May 2021

DOI: 10.1002/met.2004

RESEARCH ARTICLE

Meteorological Applications
Science and Technology for Weather and Climate

The Italian open data meteorological portal: MISTRAL

Michele Bottazzi[1] | Gabriella Scipione[1] | Gian Franco Marras[1] | Giuseppe Trotta[1] | Mattia D'Antonio[1] | Beatrice Chiavarini[1] | Cinzia Caroli[1] | Margherita Montanari[1] | Sanzio Bassini[1] | Estíbaliz Gascón[2] | Timothy Hewson[2] | Andrea Montani[2] | Davide Cesari[3] | Enrico Minguzzi[3] | Tiziana Paccagnella[3] | Renata Pelosini[4] | Paolo Bertolotto[4] | Luca Monaco[4,5] | Martina Forconi[6] | Luca Giovannini[6] | Carlo Cacciamani[7] | Luca Delli Passeri[7] | Andrea Pieralice[7]

[1]CINECA National Supercomputing Center, Casalecchio di Reno, Bologna, Italy

[2]European Centre for Medium-Range Weather Forecasts, Reading, UK

[3]ARPAe, Emilia Romagna Regional Agency for Prevention, Environment and Energy, Bologna, Italy

[4]ARPAP, Piedmont Regional Agency for Prevention, Environment and Energy, Torino, Italy

[5]Department of General Physics "A. Avogadro", University of Torino, Torino, Italy

[6]Dedagroup, Trento, Italy

[7]Civil Protection Department, Rome, Italy

Correspondence
Michele Bottazzi, CINECA, Bologna, Italy.
Email: m.bottazzi@cineca.it

Abstract

At the national level, in Italy, observational and forecast data are collected by various public bodies and are often kept in various small, heterogeneous and non-interoperable repositories, released under different licenses, thus limiting the usability for external users. In this context, MISTRAL (the Meteo Italian SupercompuTing PoRtAL) was launched as the first Italian meteorological open data portal, with the aim of promoting the reuse of meteorological data sets available at national level coverage. The MISTRAL portal provides (and archives) meteorological data from various observation networks, both public and private, and forecast data that are generated and post-processed within the Consortium for Small-scale Modeling-Limited Area Model Italia (COSMO-LAMI) agreement using high performance computing (HPC) facilities. Also incorporated is the Italy Flash Flood use case, implemented with the collaboration of European Centre for Medium-Range Weather Forecasts (ECMWF), which exploits cutting edge advances in HPC-based post-processing of ensemble precipitation forecasts, for different model resolutions, and applies those to deliver novel blended-resolution forecasts specifically for Italy. Finally, in addition to providing architectures for the acquisition and display of observational data, MISTRAL also delivers an interactive system for visualizing forecast data of different resolutions as superimposed multi-layer maps.

KEYWORDS

forecast, HPC, meteorological, MISTRAL, observation, open data

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
© 2021 The Authors. Meteorological Applications published by John Wiley & Sons Ltd on behalf of the Royal Meteorological Society.

1 | INTRODUCTION

The atmosphere is a very complex and chaotic system that is difficult to simulate. Indeed, in order to deliver forecast accuracy, particularly for longer lead times, a substantial amount of observational data is required, in addition to high-quality numerical models (Ramage, 1993).

Meteorological forecasts start out with observational data, and collecting data from various different sources, like in-situ and remotely sensed, helps to describe the multi-faceted aspects of the atmosphere, while different types of data represent different scales of phenomena (Orlanski, 1975; Rodell et al., 2004). In order to have a comprehensive view of the atmosphere, it is important to use as many data sources as possible, organized and integrated in a homogeneous way, to have a reliable forecast system upon and to show the outputs in a user-friendly visualization tool (Rautenhaus et al., 2017).

National Weather Services (NWSs) are national agencies whose main task is to provide weather forecasts and warnings of high-impact weather, by working with weather-related forecast and observational products; the main goal of these agencies is to protect and keep the society informed. To achieve this, forecasters require robust platforms that collate together wide-ranging observational and forecast data, representing different temporal and spatial scales (Glahn & Ruth, 2003). However, Italy is one of the few European countries that has no National Meteorological Service for the public. The role of National Meteorological Service is in fact fulfilled by the Italian Military Air Force (AM – Aeronautica Militare) Meteorological Service, while, at the same time, other public regional structures, mainly Agencies for Environmental Protection (ARPA – Agenzia regionale per la protezione ambientale), have started to take charge of various meteorology-related tasks.

The main actors in the Italian meteorological community are as follows:

1. Meteorological Service of AM that produces forecasts of weather and oceanic conditions at national scale, while providing also meteorological information for civil and military aviation and also representing Italy in international institutions like the World Meteorological Organization (WMO), the European Centre for Medium-Range Weather Forecasts (ECMWF) and others.
2. Civil Protection Department (DPC) that includes a central office in Rome and a network of 21 regional departments ('Functional Centres'). Together, these have the responsibility for real-time monitoring and forecasting of hazardous meteorological and hydrological conditions. Early warnings are issued to local authorities, and direct support is provided during high-impact events.
3. The network of ARPAs that build and manage the bulk of the meteorological observational networks, including both standard weather stations and sites with more sophisticated remote sensing capabilities (e.g. meteorological radar). They also produce weather forecasts at regional scale, and in most cases host the Civil Protection 'Functional Centre' for their region. Some also run operational numerical models and provide services at national scale.
4. CINECA is an inter-university consortium and the largest Italian supercomputing centre. CINECA has been a partner of ARPAE Emilia-Romagna since 1993 in its activities relating to numerical meteorological modelling, and supports DPC as a supercomputing centre for national meteorological forecasts.
5. Several private companies provide meteorological services, ranging from real-time weather forecasts to customized services for business activities affected by meteorological conditions. Some companies manage their own monitoring network and some run their own numerical models.
6. An increasing number of non-professional societies and individuals make and share observations of meteorological parameters. Data quality is very variable, but volumes are increasing.

Italy hosts a large number of ground-based meteorological networks, but it has always been very difficult to gather all data from these. AM manages the Italian synoptic network, which contributes to the international Global Telecommunication System (GTS) network: these stations have long time series and follow WMO standards for both instrumentation and data dissemination, but their number is relatively low (about 100 stations nationwide). The networks managed by the ARPAs are much more comprehensive (several thousand stations in total), and represent the bulk of the real-time data used by the Civil Protection to monitor hazardous conditions. Several other public and private companies have their own meteorological stations, and good quality observations are becoming increasingly affordable and widespread among non-professionals. All these networks are designed and managed for different purposes, and there are important technical differences between them: type and location of the instruments, frequency of measurement, data format and transmission. Data are stored in different repositories, which are often not easy to access by external users. Most ARPAs already allow free access to their real-time and historical data, but the procedures for obtaining data are different and not always straightforward. Licensing rules and conditions of use vary, and indeed are not

always clearly specified. ARPAs also share their data in the framework of the DPC and 'Functional Centres' network, but this procedure is essentially focused on precipitation and hydrology. Moreover, redistribution of the shared database is not permitted, and ARPAs can only use it for their civil protection duties. Radar observations are produced by ARPAs and Functional Centres, under the coordination of DPC, and cover almost all of Italy. DPC also manages the gathering and merging of these data: real-time maps are publicly available on DPC and ARPA websites (http://www.protezionecivile.gov.it/radar-dpc/), but numerical products and historical data are not publicly distributed. Numerical forecast models are run by different organizations in Italy. The national forecast system used for warning hazardous weather conditions (Consortium for Small-scale Modeling-Limited Area Model Italia [COSMO-LAMI]) is described in the following paragraph, but other state-of-the-art models are run operationally by ARPAs, by research institutes and by private companies. Operational modelling suites based on Moloch, Bolam and WRF models run by LAMMA Toscana, ARPA Liguria, ARPA Friuli and ISAC-CNR have been examined during the project, and they could be included in the future in the Meteo Italian SupercompuTing PoRtAL (MISTRAL) open data portal. Several numerical and graphical products based on these models are already freely available, but a common open data policy has not been defined. It has been shown that the collection of meteorological forecasts and observations in a common platform helps institutions and citizens to better anticipate and pro-actively respond to threat of extreme, high-impact weather and climate events (Abily et al., 2020; Glahn & Ruth, 2003; van Den Hurk et al., 2016). In this highly fragmented context, the MISTRAL project aims to implement a national meteorological open data portal to preserve, share, process and foster reuse of meteorological datasets (ground stations, meteorological radars and numerical forecasts) at national scale, and provide added-value services by using high performance computing (HPC) resources, thereby fuelling new business opportunities. In the present study, the different components and outcomes of the MISTRAL project and the related Web platform will be illustrated, in order to highlight the added-value and socio-economic benefits for future users. Facilitating access to observed and forecast meteorological data has been both a technical and societal challenge. Technically, there is a clear need for interoperability and harmonization of data, infrastructures and services; while from the societal standpoint, open data policies have to be put into practice. The present study is organized as follows: Section 2 describes everything related to the meteorological data offered in the MISTRAL platform, which divided into three categories: ground station, forecast and radar data. It also includes the challenges related to data licenses and data ingestion and harmonization, pointing out the necessity of HPC resources during the development of this meteorological platform. In Section 3, the main outputs from MISTRAL are illustrated through the Italy Flash Flood case study (a post-processing tool developed by the ECMWF to provide a better guidance for Flash Flood forecast), the Multimodel SuperEnsemble technique and the Meteo-Hub website visualization tools. The paper is concluded with Section 4, which summarizes the most important aspects of MISTRAL project.

2 | MISTRAL PLATFORM, ARCHITECTURE AND DATA

MISTRAL platform is based on a microservice architecture: the individual components are deployed independently and operate in a broader design. The architecture of the MISTRAL platform and its components is presented in Figure 1. The national meteorological open data portal is the main access point for citizens, public administrations and private organizations to meteorological data. The website provides useful informative sections as well as links to the underlying services.

MISTRAL catalogue is compatible with DCAT-AP_IT, the Italian extension of Data Catalog Vocabulary Application Profile (DCAT-AP) for the description of public sector catalogues, aimed at ensuring semantic interoperability among European open data portals. The catalogue exposes the metadata to other catalogues using Resource Description Framework (RDF) documents serialized according to the DCAT Italian Application Profile, in order to allow the direct upload of the managed datasets into the Italian National Data Open Portal (https://dati.gov.it, last access: 29 April 2021) and consequently into the EU Open Data Portal (https://www.europeandataportal.eu/en, last access: 29 April 2021). This service leverages the CKAN Open Data product and its capabilities that make data easily discoverable and presentable.

The core task of the ingestor is obviously to feed the dataset repository by preserving the integrity of the data. Currently, it only affects the observed data with the prospect of adapting the ingestion flow also to forecast data and to a wider stream of different data formats and protocols.

Meteo-Hub has been developed for MISTRAL and is an interface to ingested data based on two main components: a Web interface written with the Angular framework and a set of RESTful HTTP APIs. The web interface allows users to create and download personal weather data collections chosen from various forecast models, weather stations, parameters and time periods. The

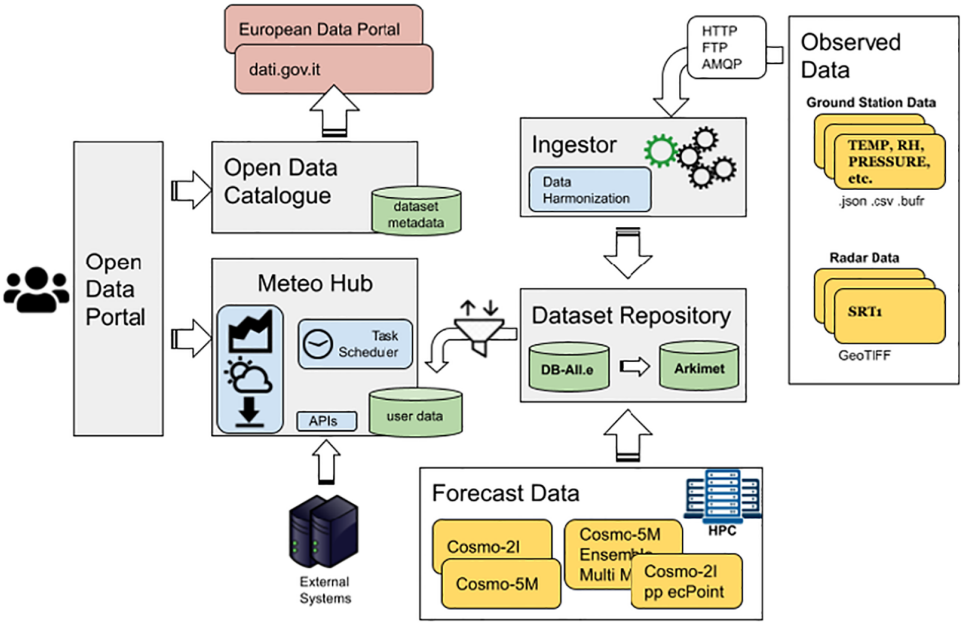

FIGURE 1 High-level representation of the MISTRAL architecture and its components

download is available in BUFR and JSON format for observed data and in GRIB format for forecast data. Different post-processing operations can also be applied to the selected data such as cumulations, averages and other statistical elaborations over time, as well as geographical elaborations, interpolations or area cropping. Users can manage their requests from their own location and are free to create as many requests as they want, until the user storage quota is reached. In addition, once the request has been raised, it can be also scheduled in order to be automatically repeated on the basis of time intervals defined by the user. When the result is ready, the user is promptly notified.

2.1 | Data description

The MISTRAL platform collects three types of meteorological data: ground station observational data collected by 11 Italian regions (Calabria, Campania, Emilia Romagna, Lazio, Liguria, Marche, Piedmont, Sardinia, Sicily, Umbria and Veneto) and the autonomous province of Bolzano, the radar data covering the whole country, the forecasts for Italy and the Mediterranean area, respectively, at 2.2 and 5 km of resolution. Moreover MISTRAL made a collaboration with Meteonetwork (https://www.meteonetwork.it/, last access: 29 April 2021), the largest Italian amateur network of meteorological ground stations, with the aim of integrating their weather observation data into the Meteo-Hub platform.

The logical data model adopted in MISTRAL has been formalized in UML and is based on the following standards:

- ISO 19156:2011, Geographic information – Observations and measurements (https://www.iso.org/standard/32574.html, last access: 29 April 2021).
- OGC/IS 08-094r1 SWE Common Data Model (https://www.opengeospatial.org/standards/swecommon, last access: 29 April 2021).
- OGC/IS 15-078r6 SensorThings API (https://www.opengeospatial.org/standards/sensorthings, last access: 29 April 2021).
- Guidelines On Data Modeling For WMO Codes (https://wis.wmo.int/DataModel, last access: 29 April 2021).

MISTRAL Data Model is based on the ISO/OGC Observation and Measurement (O&M) model [OGC 10-004r3 and ISO 19156:2011], which defines how to exchange information, describing observation acts, their results, as well as the feature involved in sampling when making observations.

Meteorological data from ground stations are encoded and archived in MISTRAL platform in BUFR format, standardized and maintained by WMO. Observations have been provided by the regions through DPC or through the own software and are in proprietary formats (CSV or JSON).

The radar data are provided by the Radar-DPC platform (http://www.protezionecivile.gov.it/radar-dpc/#/pages/dashboard, last access: 29 April 2021). Among the various data products available from DPC, MISTRAL platform ingests just Surface Rainfall Intensity (SRI) product, which identifies the areas where significant rainfall phenomena are underway. The GeoTIFF SRI product is downloaded as soon as data are available

TABLE 1 List of observed and forecast datasets released in MISTRAL and related associated licenses

Regions	Variables	Format
Calabria, Campania, Lazio, Liguria, Marche, Piedmont, Sardinia, Sicily, Umbria, Veneto, autonomous province of Bolzano	Temperature, relative humidity, wind speed, wind direction, total precipitation	Input: CSV Output: Bufr, JSON
Emilia-Romagna	Temperature, relative humidity, wind speed, wind direction, total precipitation, atmospheric pressure, snow depth, runoff, solar radiation, etc.	Input: BUFR Output: Bufr, JSON

Model	Resolution	Area	Daily run	Forecast	Format
COSMO-5M	5 km	Mediterranean	00 and 12 UTC	72 h	Input: grib Output: grib
COSMO-2I	2. km	Italy	00 and 12 UTC	48 h	Input: grib Output: grib
Italy Flash Flood	2.2 km	Italy	00 UTC	48 h	Input: grib Output: grib
Multimodel SuperEnsemble	Station coordinates	Italy	09 UTC	72 h	Input: jsonline Output: bufr

	Dataset	License
Forecast	Italy Flash Flood	cc BY 4.0
	Multimodel SuperEnsemble	CC BY 4.0
	Cosmo 5M	Copyright License
	Cosmo 2I	Copyright License
Observed data	Station data – DPC	CC BY 4.0
	Station data Emilia-Romagna	CC BY 4.0
	Station data Meteonetwork	CC BY 4.0
Maps	Forecast	CC BY-ND 4.0
	Italy Flash Flood	CC BY 4.0
	Observation	CC BY 4.0
	Multi-Layer Map	CC BY-ND 4.0
Radar	SRI – Radar-DPC	CC BY-SA 4.0

(data are updated every 5 min) and then converted to GRIB and archived in MISTRAL platform. The forecast data are available in terms of grid fields, probabilistic products or punctual time series coming from the chain of Italian operational model forecasts and post-processing procedures.

An exhaustive list of the products present in MISTRAL is shown in Table 1.

2.2 | Ingestion, harmonization and dataset repository

The ingestion and harmonization of observational data are coordinated in MISTRAL by Apache NiFi (https://nifi.apache.org/, last access: 29 April 2021), an open source application for the design and management of data transformation flows, and relies on two different operational databases: DB-All.e (https://github.com/ARPA-SIMC/dballe, last access: 29 April 2021) and Arkimet (https://github.com/ARPA-SIMC/arkimet, last access: 29 April 2021). The ingestion component relies on RabbitMQ to offer an AMQP endpoint to data providers and also relies on a PostgreSQL DB to persist operational information such as the list of products to be downloaded, the timestamp of the various operations, their result and the polling period. This component offers a unique control point for various data flows, monitoring, easiness of data flow updating, scalability and reliability.

DBALL.e is a fast, on-disk database used exclusively to store observed data. It allows to manage large amounts of data using its simple API, and provides tools to import and export in the standard BUFR format. The main use of DBALL.e is to accommodate and make readily available the most current data while the older ones are periodically archived. Arkimet plays the role of an actual archive. Here, both forecast and observed data are collected and stored for a long time. It currently supports data in GRIB, BUFR, HDF5 and VM2 formats. Arkimet manages a set of datasets, each of which contains homogeneous data stored in segments. It exploits the commonalities among the data in a dataset to implement a fast,

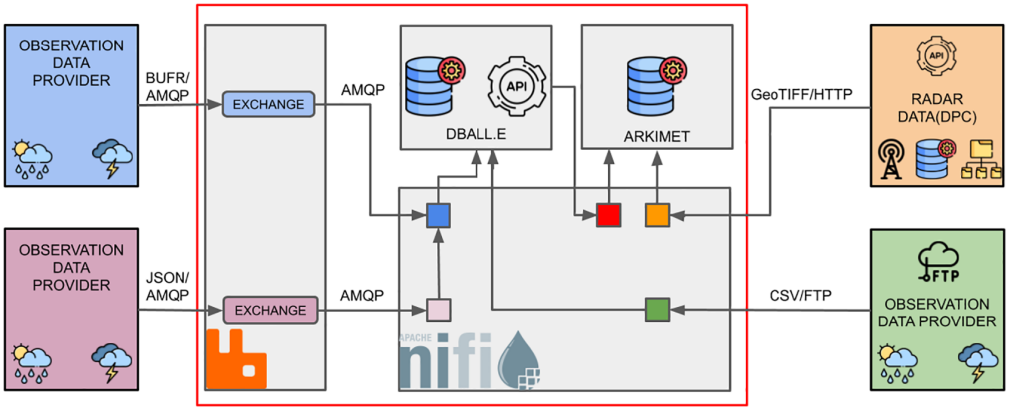

FIGURE 2 Schema of NiFi flows for ingestion and harmonization

powerful and space-efficient indexing system. When data are ingested into Arkimet, it is scanned and annotated with metadata, such as reference time and product information, and then it is dispatched to one of the datasets configured in the system.

Data from ground stations are currently provided in two different ways: made available and retrieved from an FTP server or directly transferred by data provider via AMQP protocol. Differently, the radar data are harvested by querying a Web service. The ingestion process is coordinated by an orchestrator, which channels the data into the right pipeline and ensures that the data are properly stored. This process, also called harmonization, requires interpretation of the data and may involve transformations and normalizations for quality assurance of the data.

The four main flows developed in MISTRAL are presented in Figure 2: the radar data are ingested into Arkimet, and the ground station data into DBAll.e via AMQP (pink/blue flow) and FTP (green flow).

Radar data are ingested into Arkimet. Several radar data products are available on the DPC platform (WIDE – Weather Ingestion Data Engine) from which they can be downloaded by means of the platform's APIs. The combined use of the APIs enables a remote client to search for a certain product of a specific time and then download it. The formats of the downloadable data are raster GeoTIFF and vector shapefiles. NiFi checks for the presence of updated releases of the SRI product at regular frequency (data polling).

Observed data are stored in the short-term repository (DBAll.e) to speed-up retrieving the most recent data by MISTRAL users. Ground station data are delivered by providers queuing BUFR or JSON messages in an AMQP endpoint exposed by MISTRAL. To implement the AMQP channels, MISTRAL deploys a message broker component implemented by RabbitMQ (https://www.rabbitmq.com/, last access: 29 April 2021) where 'exchanges' and 'queues' are defined respectively to receive messages from providers and to publish them for NiFi. RabbitMQ provides a robust implementation of AMQP protocol, so the architecture of the ingestion module includes this component for exposing the exchange endpoints. Ground station data are provided by the DPC via FTP in the form of CSV-formatted files and is stored into DBAll.e after transformation to BUFR. DPC uploads a new CSV file every 20 min, containing the available observations for the previous 2 h. DPC also provides a second type of CSV file describing the sensor properties for the administrative regions that authorized the project to publish the data from their sensor networks (10 regions out of 21 at the time of writing this paper). The measurements currently provided by DPC are limited to certain sensor types: thermometer, pluviometer, hygrometer, barometer and anemometer. Older data are transferred daily from the short-term repository DBAll.e to the long-repository of Arkimet. This transfer is achieved with the support of a temporary DB, whose aim is to support the flow management and to log its results (either success or errors).

2.3 | COSMO-LAMI agreement and weather forecasting activities

MISTRAL portal receives as input the meteorological forecast data generated in the national agreement called LAMI, signed by the Meteorological Office of the Italian Air Force, ARPAE and ARPAP, that regulates the joint effort to develop, and operationally manage, the national numerical forecast system in Italy. The data are available in terms of grid fields, probabilistic products (such as rain forecasts for the forecast of floods) or punctual time series coming from the modelling chain of Italian operational forecasts and post-processing fields (such as probability of thunderstorm). The Department of National Civil Protection has assigned to the signatories of the LAMI agreement the task of systematically executing a suite of state-of-the-art numerical forecasting models to the best of their knowledge and to provide the results in graphic

and numerical forms to the Department and to Civil Protection centres in Italy. The configuration of the model and the modelling suites is decided by the LAMI partners; ARPAE Idro-Meteo-Climate service is responsible for the creation and management of part of the LAMI suites; ARPAE has assigned the CINECA supercomputing center, the task of execution and monitoring of part of the LAMI Numerical Weather Prediction (NWP) suites. LAMI NWP suites are currently based on the COSMO model and include the two series of models: *COSMO-5M* and *COSMO-2I*. *COSMO-5M* performs a forecast on the Mediterranean area up to 72 h with a 5 km grid step, started twice a day by AM analysis and using the Integrated Forecast System (IFS) boundary conditions of the ECMWF.

COSMO-2I performs analysis and forecasts over the Italian area with a 2.2 km grid step. It is divided into four main series:

1. COSMO-2I-assim performs almost continuous data assimilation using the data assimilation system of COSMO's KENDA ensemble. It includes 40 members of the ensemble and runs in 1 h forecast-update phases, using the COSMO ensemble of AM and COSMO-5M as boundary conditions;
2. COSMO-2I-fcast performs a forecast up to 48 h, initialized twice a day by COSMO-2I-assim and using the COSMO-5M boundary conditions;
3. COSMO-2I-fcruc (RUC Rapidly Updating Cycle) performs a forecast of up to 18 h, initialized eight times a day by COSMO-2I-assim and using COSMO-5M boundary conditions, for short-range forecasting and casting;
4. COSMO-2I-fcens performs a daily ensemble forecast with 20 ensemble members, up to 51 h, initialized by COSMO-2I-assim and using AM's COSMO ensemble as boundary conditions.

2.4 | Data license

One of the goals of the project is to facilitate and promote the reuse of meteorological data by providing free access to both observational and forecast data as well as visualization tools products. Data collections integrated into the platform are covered in Europe by '*sui generis*' right of the database maker established by Directive 96/9/EC instead of classical copyright. This law covers operations of extraction and reuse of substantial parts of databases for which a significant investment was necessary. To manage the licenses assigned to the works derived from the applications developed in MISTRAL, for example, the data visualization maps, a data model has been designed. In this data model only data with compatible license are integrated, creating the conditions to manage their use and redistribution without incurring problems of license incompatibility. For this reason, where possible, it has been chosen to distribute the data under a CC BY 4.0 license (https://creativecommons.org/licenses/by/4.0/deed.en, last access: 29 April 2021), which takes this right into consideration. In the case of open licenses already defined, as for the civil protection RADAR data, in MISTRAL these are collected and redistributed with the same type of license: CC BY-NC-SA 3.0 (https://creativecommons.org/licenses/by-nc-sa/3.0/deed.en, last access: 29 April 2021). For the forecast data, the MISTRAL project is authorized to use the forecast data of the COSMO-LAMI agreement (ARPAE-ARPAP-AM) because two of the project partners, ARPAE and ARPAP, are part of the agreement and AM has been available. The original software packages integrated inside the MISTRAL project are the following: (1) DB-All.e (https://github.com/ARPA-SIMC/dballe, last access: 29 April 2021): software for the management of observational data; (2) Arkimet (https://github.com/ARPA-SIMC/arkimet, last access: 29 April 2021): software for archiving modelling and observational data; (3) libsim (https://github.com/ARPA-SIMC/libsim/, last access: 29 April 2021): software for data post-processing. All three are licensed under the GNU General Public License version 2.0.

2.5 | HPC usage: Creating and upgrading forecasts

The operational execution of Numerical Weather Prediction (NWP) models, as LAMI suites, requires large amounts of computational and high-quality I/O resources to produce results in reasonable time intervals, in the order of a few hours. Furthermore, the demand for more accurate weather forecasts, together with the scientific development of atmospheric models, bring to an increase in the complexity of model physical parameterizations and to an increase of resolution, thus implying an increase in the computing, storage and networking resources needed for managing the simulations. For this reason, HPC has become fundamental in meteorology mainly in the accomplishment of numerical simulations for weather forecasts and climate investigation. Moreover, with the appearance of ensemble predictions, the amount of forecast data grows by an order of magnitude compared with deterministic forecast method, and the need for supercomputing arises not only in the phase of integrating the model equations, but also in the successive phase of post-processing the results of the ensemble.

The supercomputing resources are provided to the MISTRAL platform by the partner CINECA, the Italian

TABLE 2 HPC resources required to run the COSMO models

COSMO-5M	Galileo	Meucci
Forecast	30 nodes	39 nodes
	960 CPU-cores	1248 CPU-cores
COSMO-2I	**Galileo**	**Meucci**
Forecast	30 nodes	39 nodes
	960 CPU-cores	1248 CPU-cores
Ensemble forecast	27 nodes	–
	864 CPU-cores	
Ensemble data assimilation	25 nodes	25 nodes
	800 CPU-cores	800 CPU-cores

Notes: Both the HPC systems, Galileo and Meucci, are hosted by CINECA.

national facility for supercomputing applications and research and one of the largest infrastructures in Europe (https://www.top500.org/lists/top500/2020/06/, last access: 29 April 2021). CINECA's expertise in the area of supercomputing for meteorological forecasts is long-standing. For over 30 years, CINECA has been providing the computing and human resources for implementing, running and monitoring part of the LAMI NWP suites. The final data are made available within 1 h from receiving the input data. In particular, the forecast suites COSMO-5M and COSMO-2I are run by CINECA on two different HPC clusters (Galileo and Meucci) several times a day in operational mode. The number of computing nodes required for the accomplishment of these model runs in operational times is shown in Table 2.

The operational production of numerical models for weather simulation is a highly critical service (unattended mode 365 days a year, strict timing of data availability, activation of the service when input data are available), and in order to efficiently provide this type of service, the entire infrastructure must be extremely reliable and redundant in each component, suitable for a stable production environment. The HPC has been mandatory also for the Flash Flood use case that was implemented on the MISTRAL platform with the collaboration of the partner ECMWF (Centre of Excellence European Centre for Medium-Range Weather Forecast). In the ECMWF forecast post-processing, a new 100 member ensemble is generated for each of the 51 ensemble members for each lead time interval. Due to the computation demand of the post-processing, HPC resources were used. The computational demands this creates are very high, and in tests at ECMWF they found that the only viable option for operational production was to use HPC resources.

3 | OUTPUTS FROM MISTRAL

3.1 | Multimodel SuperEnsemble technique for forecasting temperature and humidity

Temperature and humidity forecasts for Italian weather stations included in the Regional Civil Protection Functional Centers network are provided within the MISTRAL project, as part of the ARPAP collaboration. The Multimodel SuperEnsemble technique is a powerful post-processing method for the estimation of some weather forecast parameters with the aim of reducing direct model output errors. This type of forecast arose more than 20 years ago (Krishnamurti et al., 1999) and uses multiple regression techniques in order to determine the coefficients from a set of both NWP forecasts and observations from a dense network of weather stations. The coefficients derived from the regression are used in the superensemble calculation. This technique requires several adequately weighted model outputs, where weights are calculated during the so-called training period (Krishnamurti et al., 1999, 2000; Williford et al., 2002). The Multimodel SuperEnsemble technique, now applied in many meteo-climatic research and operational branches (Krishnamurti et al., 2016), has been operationally implemented at ARPAP in 2015 (Cane & Milelli, 2005). Over the years, it has undergone continuous improvements, including software modifications, adjustments to the number of models, to the training period and to the quality of the observed data. Quoting Hagedorn et al. (2005) stated 'the key to the success of the multimodel concept lies in combining independent and skilful models, each with its own strengths and weaknesses'. Over the years several studies, also carried out on a local scale (Cane & Milelli, 2006; Weigel et al., 2008), have shown that Multimodel SuperEnsemble forecasts perform better than forecasts of both single models and other post-processing techniques (Kalman filtering, poor man's ensemble, non-weighted multimodel). The importance of implementing this technique within the MISTRAL project is based on two factors, which are firstly, the need to have solid and reliable predictions of ground meteorological parameters, and secondly, the flexibility of the technique itself: in fact, in order to improve the accuracy of the weights and consequently the reliability of multimodel forecast, it is possible to use both new models, from a forecast perspective, and new

weather stations of different networks, from an observational perspective.

3.2 | Italy flash flood

The Italy Flash Flood use case was implemented with the collaboration of ECMWF, which developed an application for Italy based on recent advances in HPC-based post-processing of global ensemble forecasts (ECMWF ENS IFS), specifically forecasts of rainfall. This application was created by combining the post-processed output with rainfall output provided by the Italian high-resolution 2.2 km COSMO ensemble (COSMO-2I-EPS), to which the main goal is improving Flash Flood prediction for that part of the Mediterranean area. This use case is a demonstration of the use of supercomputer architecture to make real-time product delivery tractable. The final Flash Flood use case is divided into three main computation steps:

1. The Point-Rainfall product aims to deliver probabilities for point measurements of rainfall (within a forecast model gridbox). 'ecPoint' is the name given to the post-processing philosophy and software package. The first task involved porting pre-existing ECMWF Point-Rainfall code to the new national supercomputer portal at CINECA, and adapting and optimizing the code to fit the specific requirements of the new platform, such as its filesystem, I/O constraints and memory availability. ECMWF Point-Rainfall forecast production, for 12-h precipitation totals, is running on an operational basis on the ECMWF supercomputer in Reading (UK). The development, testing and verification processes for 6-hourly Point-Rainfall was carried out in MISTRAL project, but the computational requirements increased twice compared with the 12-h Point-Rainfall production. The techniques and scientific aspects applied to develop 6-h Point-Rainfall were based on Hewson and Pillosu (2020). The Point-Rainfall distribution forecasts are calculated for every gridbox (with 18 km horizontal resolution) in the world for each of the 51 ensemble member and for each time interval (65 intervals), and it generates 100 new 'calibrated' ensemble members per each of the 51 original ensemble members, so 5100 ensemble members in total. This aspect of the post-processing procedure becomes computationally challenging, especially for an operational product and for producing the data dissemination on time. Using the parallelization capabilities of the Galileo system in CINECA HPC, the computational time for the whole process is less than 40 min, which could be up to 42 CPU hours of computation. Four different tasks are running per time step, two of them required parallelization. This implies the overall generation of 5100 ensemble members for each of 63-time windows, which amounts to creating 5 TB of data (for just the rainfall totals) at an intermediate stage. The computational demands are very high, and in tests at ECMWF, it has been found that the only viable option for operational production is to use HPC resources. Parallel processes are used in two main steps of the post-processing: in the first one, which calculates the weather types for each grid point, we used 2 nodes with 25 and 26 CPU per node, respectively, running at the same time this task for each of the 51 ECMWF ensemble members. Then, the calculation of the percentiles in each grid point based on 5100 ensemble members (generated in the previous step) is computed and due to the huge amount of data generated, the computation is divided into 10 subareas, which will run in parallel in 1 node and 10 CPUs.
2. The second task involved post-processing of a second input source, specifically the recently developed COSMO-2I-EPS. The variable-size neighbourhood post-processing described in Dey, Roberts, et al. (2016b) and Dey, Plant, et al. (2016a) was adapted and applied to COSMO-2I-EPS and the final post-processing product was called COSMO-2Ipp. Based on a neighbourhood approach, the scales over which ensemble members reach a specified level of agreement (S) were calculated, at each grid point in the domain, to give a measure of the location-dependent believable scales for an ensemble forecast, and it corresponds to the scale-agreement level where the ensemble members become sufficiently similar to provide a useful, trustworthy forecast.
3. The final stage is to produce the final product, called COSMO-2Ipp-ecPoint, and deliver a fully operational real-time system for forecasting rainfall probabilistically, based on the model output blending of both post-processing products. COSMO-2Ipp was combined with the ECMWF 6 h Point-Rainfall output in such a way that the most skilful aspects of the two systems can be best exploited to provide products for the end users. Compared with using raw model output, we deliver gains across the full range of rainfall severity, from small totals right through to the extreme amounts, which can lead to devastating flash floods. COSMO-2Ipp-ecPoint is a set of probabilistic 6-h precipitation forecast products, for Italy and surrounding countries, up to day 10. It consists of two different types of product, depending on the lead time of the forecast: From 0 to 48 h, the blending of 6 h Point-Rainfall and COSMO-2pp is produced with different weights applied in that blending, depending on the lead time. More weight is given to COSMO-2pp at

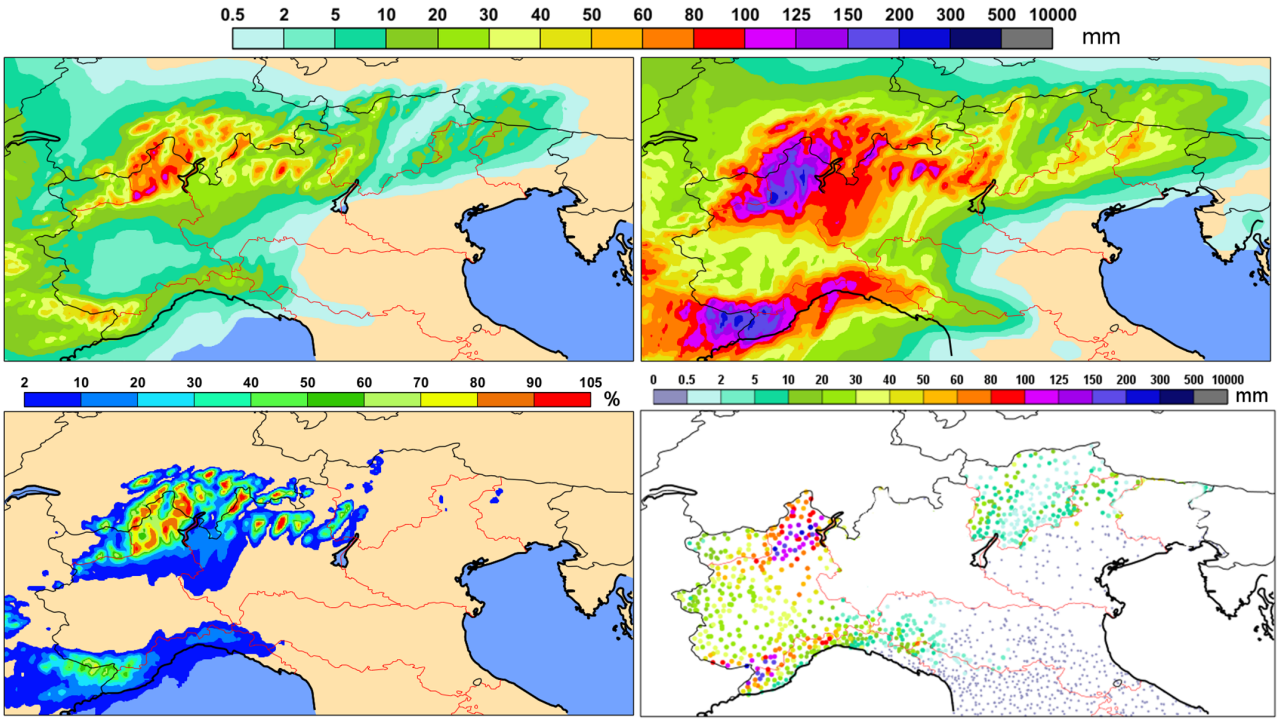

FIGURE 3 Italian Flash Floods products example with initialization on 2 October 2020 at 00 UTC and valid at 21 UTC the same day. The plots represent the 50th percentile (top left), the 99th percentile (top right), the probabilities of 6 h precipitation greater than 50 mm (bottom left) and the corresponding observations available in the MISTRAL portal for the same period (bottom right)

shorter lead times, tapering to about zero at 48 h. The horizontal resolution of the product is 2.2 km and time resolution is 3 h. Then, from 48 to 240 h (day 10), only 6 h Point-Rainfall is available, at 18 km horizontal resolution. The time resolution is 3-hourly from 48 to 144 h and 6-hourly from 144h to 240 h. The 'tapered blending' strategy means that through all the lead times, and indeed across the 48 h 'barrier', the forecast evolution should look relatively smooth to users, with more detail apparent at the shortest leads. The intention was to create an accurate, useful, seamless forecast. The final products are created twice a day, combining firstly 6 h Point-Rainfall from the 12 UTC ECMWF IFS ENS with the 21 UTC COSMO-2I-EPS post-processing outputs, and then when the 00 UTC ECMWF ensemble data become available, that is, combined with the same 21 UTC COSMO data, to deliver an 'update' of the products. The final products are produced as probabilities of exceeding 6 h precipitation thresholds (5, 10, 20 and 50 mm) and representing the percentiles 1, 10, 25, 50, 70, 75, 80, 90, 95 and 99. EcPoint is specifically designed to improve the reliability and discrimination ability of the forecast. In verification, it shows some particularly good results for large totals. Furthermore, by combining with the COSMO ensemble output, we further improve the forecasts. And in certain situations, we can give increased specificity regarding where the most extreme totals are most likely to occur. Figure 3 shows some examples of the Italian Flash Flood products valid for the 2nd October 2020 at 21 UTC. More than 200 mm in 6 h (and more than 600 mm in 24 h) were registered in some areas in the Northwestern part of Italy, creating devastating localized landslides and flash floods. Both products, the percentiles (50th and 99th in this case) and the probabilities of precipitation greater than 50 mm in 6 h, show a good agreement with the observations for the same period of time. Note that the observations plotted here are only in the regions where a data agreement with MISTRAL exists.

The use of HPC resources is crucial for this use case, because large amounts of computational and high-quality I/O resources are required for operational production of numerical model weather forecasts. To provide predictions with great accuracy, the meteorological models have high-resolution grids and need to be run with short time steps. This implies a need for elevated supercomputing power and performance. On the other side, HPC is also necessary for ensemble forecasting, where the meteorological model is executed several times starting from slightly different initial conditions. In this use case, the ensemble forecasts will be transformed into much more useful information, in real time, by using the

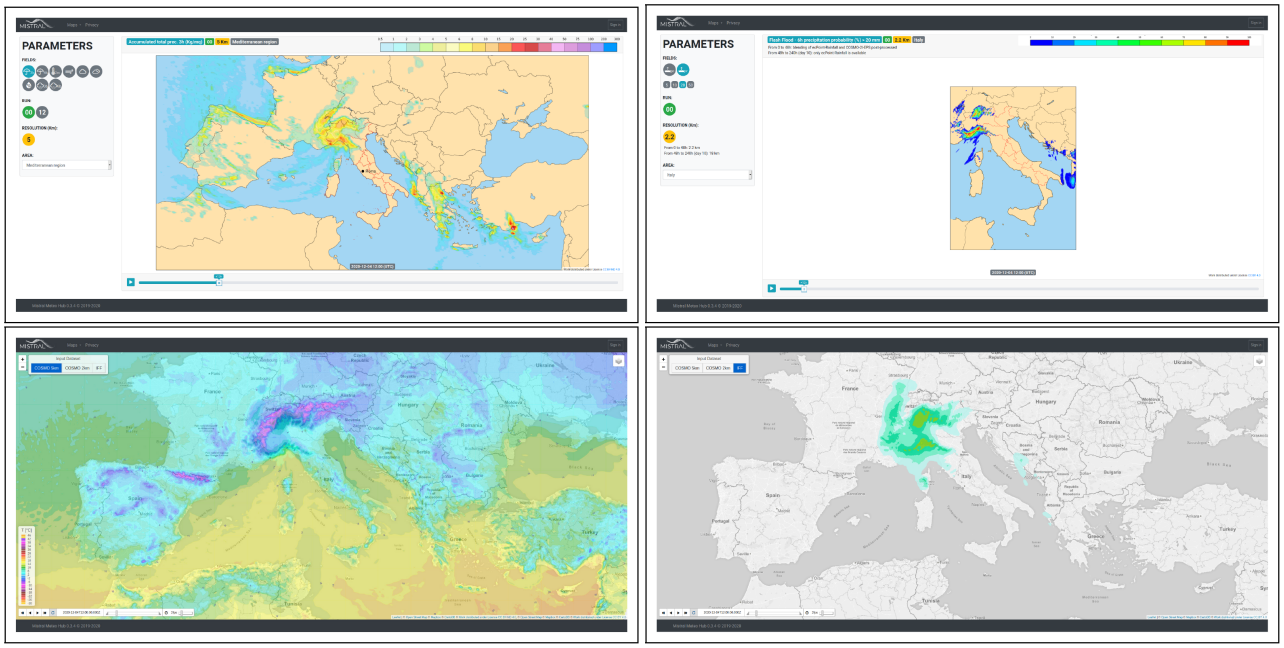

FIGURE 4 Visualization of different products for 4 December 2020 at 12 UTC. The static representation of 6 h accumulated total precipitation for COSMO-5M (top left) and of the Italy Flash Floods 6 h percentage probability of a precipitation ≥20 mm (top right). The multi-layer representation of the COSMO-5M temperature at 2 m height (bottom left) and the Italy Flash Floods 6 h precipitation at 20th percentiles (bottom right)

supercomputing resource to analyse and post-process the raw forecast data relatively quickly.

3.3 | Data visualization: Dynamic maps

The MISTRAL portal provides some tools for the visualization of observed and forecast data. Observed data from ground stations are displayed on a customized Web viewer developed for MISTRAL project: the measurement of the weather variables are visualized as geo-referenced markers over an Open Street Map layer. Moreover, the viewer provides the meteograms of the data for each variable.

Both COSMO forecast and Italy Flash Flood are available in the MISTRAL visualization tools for forecast data (Figure 4).

The graphical representation is produced using the Magics package (https://github.com/ecmwf/magics, last access: 29 April 2021), created and managed by ECMWF. Temperature at 2 m, rainfall and snowfall at 3 and 6 h, relative humidity, wind direction and intensity and the cloud cover at three different levels (low, medium and high) are displayed for COSMO, while percentile and percentage are displayed for Italy Flash Flood. For all data, two different kinds of visualization tools are available:

- Standard (static) map: raster map images in `.png` format are created for each meteorological variable. These images are not geo-referenced and are not possible to dynamically change the zoom level during the visualization. Is the historical Web viewer provided by CINECA to ARPAE that in the MISTRAL project has been improved and integrated in the platform;
- Multi-layer map: each meteorological variable is represented on a tiled Web map. This provides a dynamic visualization allowing navigation and zooming, as well as the superimposition of multiple layers representing different meteorological variables. It has been developed for the MISTRAL project.

We have chosen to represent the maps as a tiled Web map because tiling is a standardized process for Web map applications. A tiled Web map is a map displayed in a browser that is represented by seamlessly joining files of sub-pictures requested individually on the Internet. It is the most popular way to view and navigate maps, replacing other methods such as WMS that typically displays a single large image, with the arrow buttons to navigate in nearby areas. Not all tiles are required simultaneously; only those necessary to cover the area visible in the map viewer, in addition to the surrounding ones, are transferred from the server to the client and stored there in the cache. This method is used because the map application becomes faster and the server is not busy due to sending bulk data to a few users. The idea is to prepare, on the server side, the panels that contain the

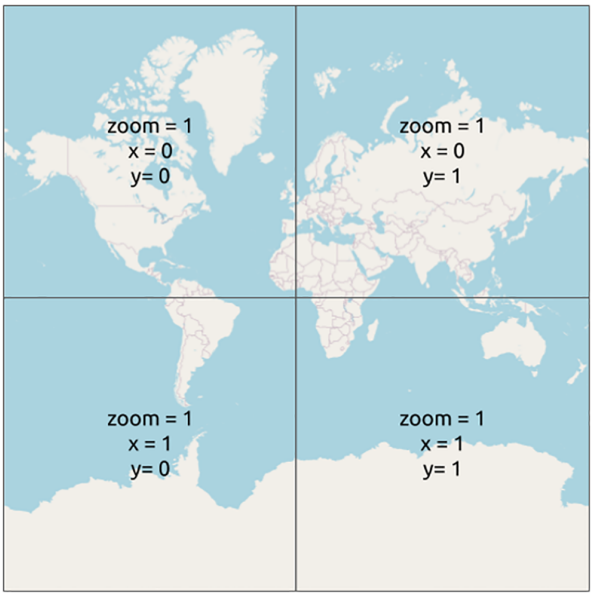

FIGURE 5 Example of OpenStreetMap (https://www.openstreetmap.org/, last access: 29 April 2021) tiled map, with zoom equal to one

data and, on the client side, to load the panels asynchronously in order to provide a better user experience (Wendlik et al., 2011). The properties of tiled Web maps that require convention or standard include the size of the tiles, the numbering of the zoom levels, the projection to be used, the way in which the individual tiles are numbered or otherwise identified and the method to request them. Web maps generally follow some conventions. The tiles are presented as images with a size of 256 × 256 pixels. At the outermost zoom level (zoom level 0), the whole world can be transformed into a single map tile where each zoom level doubles in both dimensions, so a single tile is replaced by four tiles when zooming. In MISTRAL the zoom ranges between 5 and 7 for the COSMO-5km and between 6 and 8 for COSMO-2I. The Web Mercator (EPSG: 3857) projection is used with latitude limits of approximately 85 degrees. A numbering scheme for tiles with X (horizontal) and Y (vertical) is presented in Figure 5. Images are served through a Web server, with a URL like http://.../Z/X/Y.png, where Z is the zoom level and X and Y identify the box. Each zoom level is a directory, each column is a subdirectory, and each tile in that column is a file.

4 | CONCLUSION

In the present study, we present and describe the architecture of the Meteo Italian SupercompuTing PoRtAL (MISTRAL) project: the first Italian supercomputing portal of meteorological open data. The main purpose of MISTRAL is to compensate for the lack of a national service for the management of meteorological data. These data are in fact currently collected, managed and redistributed by the individual regional agencies, in different repositories and distributed under different licenses. For this reason, MISTRAL proposes itself as a national portal in order to increase and improve the use of meteorological data not only by civil users but also by researchers. At the time of writing the article, station data from more than half of the Italian regions are collected and redistributed through the portal, in addition to data from the Meteonetwork amateur network.

Other than observed data, the forecast data developed within the Consortium for Small-scale Modeling-Limited Area Model Italia (COSMO-LAMI) agreement (COSMO-5M and COSMO-2I) are provided. To improve the reusability of the data, all the forecast data can be post-processed before downloaded, both spatially and temporally. MISTRAL also includes the Italy Flash Floods use case, implemented with the collaboration of European Centre for Medium-Range Weather Forecasts (ECMWF), which provides a real-time application for Italy and neighbouring areas based on recent advances in high performance computing (HPC)-based post-processing of precipitation forecasts.

The final products are produced as probabilities of exceeding 6 h precipitation thresholds (5, 10, 20 and 50 mm) and representing the percentiles 1, 10, 25, 50, 70, 75, 80, 90, 95 and 99. Tools for data visualization are also provided, both for observed and forecast data. An interactive method for visualizing raster data is also provided. This type of visualization is based on the system of tiles and allows a dynamic visualization of the maps and the overlapping of fields. This type of display is provided both for COSMO-5M and COSMO-2I forecast data and for Italy Flash Flood data. All these datasets and the maps are released under open licenses CC BY 4.0 and CC BY 4.0-ND in order to improve the usability. To conclude, the aim of MISTRAL is therefore not only to improve the use of the meteorological data currently present in it but to become a national reference point for meteorological datasets not yet included in the portal and to represent a platform on which a future Italian National Weather Service (NWS) can rely.

AUTHOR CONTRIBUTIONS

Michele Bottazzi: Software (equal); validation (equal); visualization (equal); writing – original draft (equal); writing – review and editing (equal). **Gabriella Scipione:** Conceptualization (equal); funding acquisition (equal); project administration (equal); supervision (equal); writing – original draft (equal). **Gian Franco Marras:** Data

curation (equal); software (equal); validation (equal); visualization (equal). **Giuseppe Trotta:** Software (equal); validation (equal); visualization (equal). **Mattiia D'Antonio:** Software (equal); validation (equal); writing – original draft (equal). **Beatrice Chiavarini:** Software (equal); validation (equal). **Cinzia Caroli:** Project administration (equal); validation (equal). **Margherita Montanari:** Data curation (equal); project administration (equal); writing – original draft (equal). **Sanzio Bassini:** Funding acquisition (equal); resources (equal). **Estibaliz Gascòn:** Formal analysis (equal); software (equal); validation (equal); writing – original draft (equal); writing – review and editing (equal). **Tim Hewson:** Formal analysis (equal); validation (equal); writing – original draft (equal); writing – review and editing (equal). **Andrea Montani:** Formal analysis (equal); validation (equal); writing – original draft (equal). **Davide Cesari:** Formal analysis (equal); software (equal); supervision (equal); writing – review and editing (equal). **Enrico Minguzzi:** Software (equal); writing – original draft (equal). **Tiziana Paccagnella:** Conceptualization (equal); funding acquisition (equal). **Renata Pelosini:** Conceptualization (equal); funding acquisition (equal); supervision (equal). **Paolo Bertolotto:** Formal analysis (equal); writing – original draft (equal). **Luca Monaco:** Formal analysis (equal); writing – original draft (equal). **Martina Forconi:** Data curation (equal); software (equal). **Luca Giovannini:** Software (equal). **Carlo Cacciamani:** Conceptualization (equal); funding acquisition (equal); supervision (equal). **Luca Delli Passeri:** Data curation (equal). **Andrea Pieralice:** Data curation (equal).

ORCID

Michele Bottazzi https://orcid.org/0000-0002-5381-8389
Cinzia Caroli https://orcid.org/0000-0002-9086-2629

REFERENCES

Abily, M., Gourbesville, P., Filho, E.D.C., Llort, X., Rebora, N., Sanchez, A. et al. (2020) Anywhere: enhancing emergency management and response to extreme weather and climate events. In: *Advances in Hydroinformatics*. Singapore: Springer, pp. 29–37.

Cane, D. & Milelli, M. (2005) Use of multimodel superensemble technique for mountain-area weather forecast in the olympic area of torino 2006. *Hrvatski meteorološki časopis*, 40(40), 236–239.

Cane, D. & Milelli, M. (2006) Weather forecasts obtained with a multimodel superensemble technique in a complex orography region. *Meteorologische Zeitschrift*, 15(2), 207–214.

van Den Hurk, B.J.J.M., Bouwer, L.M., Buontempo, C., Döscher, R., Ercin, E., Hananel, C. et al. (2016) Improving predictions and management of hydrological extremes through climate services: www.imprex.eu. *Climate Services*, 1, 6–11.

Dey, S.R.A., Plant, R.S., Roberts, N.M. & Migliorini, S. (2016a) Assessing spatial precipitation uncertainties in a convective-scale ensemble. *Quarterly Journal of the Royal Meteorological Society*, 142(701), 2935–2948.

Dey, S.R.A., Roberts, N.M., Plant, R.S. & Migliorini, S. (2016b) A new method for the characterization and verification of local spatial predictability for convective-scale ensembles. *Quarterly Journal of the Royal Meteorological Society*, 142(698), 1982–1996.

Glahn, H.R. & Ruth, D.P. (2003) The new digital forecast database of the National Weather Service. *Bulletin of the American Meteorological Society*, 84(2), 195–202.

Hagedorn, R., Doblas-Reyes, F.J. & Palmer, T.N. (2005) The rationale behind the success of multi-model ensembles in seasonal forecasting—I. Basic concept. *Tellus A: Dynamic Meteorology and Oceanography*, 57(3), 219–233.

Hewson, T.D. & Pillosu, F.M. (2020) *A new low-cost technique improves weather forecasts across the world*. arXiv preprint arXiv:2003.14397.

Krishnamurti, T.N., Kishtawal, C.M., LaRow, T.E., Bachiochi, D.R., Zhang, Z., Williford, C.E. et al. (1999) Improved weather and seasonal climate forecasts from multimodel superensemble. *Science*, 285(5433), 1548–1550.

Krishnamurti, T.N., Kishtawal, C.M., Shin, D.W. & Williford, C.E. (2000) Improving tropical precipitation forecasts from a multi-analysis superensemble. *Journal of Climate*, 13(23), 4217–4227.

Krishnamurti, T.N., Kumar, V., Simon, A., Bhardwaj, A., Ghosh, T. & Ross, R. (2016) A review of multimodel superensemble forecasting for weather, seasonal climate, and hurricanes. *Reviews of Geophysics*, 54(2), 336–377.

Orlanski, I. (1975) A rational subdivision of scales for atmospheric processes. *Bulletin of the American Meteorological Society*, 56, 527–530.

Ramage, C.S. (1993) Forecasting in meteorology. *Bulletin of the American Meteorological Society*, 74(10), 1863–1872.

Rautenhaus, M., Böttinger, M., Siemen, S., Hoffman, R., Kirby, R.M., Mirzargar, M. et al. (2017) Visualization in meteorology—a survey of techniques and tools for data analysis tasks. *IEEE Transactions on Visualization and Computer Graphics*, 24(12), 3268–3296.

Rodell, M., Houser, P.R., Jambor, U.E.A., Gottschalck, J., Mitchell, K., Meng, C.-J. et al. (2004) The global land data assimilation system. *Bulletin of the American Meteorological Society*, 85(3), 381–394.

Weigel, A.P., Liniger, M.A. & Appenzeller, C. (2008) Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts? *Quarterly Journal of the Royal Meteorological Society*, 134(630), 241–260.

Wendlik, V., Karut, I. & Behr, F.-J. (2011) *Tiling concepts and tile indexing in internet mapping APIs*. Karlsruhe, Germany: AGSE, p. 116.

Williford, C.E., Krishnamurti, T.N., Correa-Torres, R.J., Cocke, S., Christidis, Z. & Kumar, T.S.V.V. (2002) Real time multi-analysis/multimodel superensemble forecasts of precipitation using trmm and ssm/i products. *Monthly Weather Review*, 131 (8), 1878–1894.

How to cite this article: Bottazzi, M., Scipione, G., Marras, G. F., Trotta, G., D'Antonio, M., Chiavarini, B....Pieralice, A. (2021). The Italian open data meteorological portal: MISTRAL. *Meteorological Applications*, *28*(4), e2004. https://doi.org/10.1002/met.2004