European Commission

Proceedings of the 2021 conference on

Big Data from Space (BiDS'21)

---From Insights to Foresight---

18-20 May 2021

Edited by P. Soille, S. Loekken, and S. Albani

SatCen
EUROPEAN UNION SATELLITE CENTRE
Analysis for decision making

European Commission

esa
European Space Agency

Joint Research Centre

EUR 30697 EN

Contact information: Pierre.Soille at ec.europa.eu

EU Science Hub
https://ec.europa.eu/jrc

JRC125131

EUR 30697 EN

PDF ISBN 978-92-76-37661-3 ISSN 1831-9424 doi:10.2760/125905

Luxembourg: Publications Office of the European Union, 2021

How to cite this report: *Proceedings of the 2021 conference on Big Data from Space*, Soille, P., Loekken, S. and Albani, S., eds., EUR 30697 EN, Publications Office of the European Union, Luxembourg, 2021, ISBN 978-92-76-37661-3, doi:10.2760/125905, JRC125131.

Preface

While data have always been at the source of any scientific or technical endeavour, the term Big Data has gained momentum at the beginning of this millennium given the vast amounts of digital data generated from ever increasing sources ranging from individuals and the internet of things to sophisticated sensors in laboratories, on the ground, and in space. With an initial focus on addressing technical aspects such as handling data volume, velocity and variety, Big Data is now more about our capacity to extract meaningful information (insights) from multi-source data, and more recently also deriving the *foresight* to inform and support decisions that will shape our future.

In this context, *Big Data from Space* refers to the massive spatio-temporal Earth and Space observation data collected by a variety of ground-based & space-borne sensors and the synergy with data coming from other sources and communities. The continuous growth of Big Data from Space is motivated by the need for answering major societal challenges related to the impact of human activities on our planet in the case of Earth Observation, or fundamental questions related to the origin of our Universe in the case of space observation. While the growth of data volume, velocity, and variety is matched by technological advances related to sensors as well as information and communication technologies, the extraction of insights from the generated data is enabled by breakthroughs in data science and in particular thanks to recent progress in artificial intelligence. These developments are empowering new approaches and applications in various and diverse domains influencing life on Earth and societal aspects, from sensing cities, monitoring human settlements and urban areas to climate change, sustainable development goals, and security.

The main objectives of the 2021 Big Data from Space conference (BiDS'21) are:

- Bring to the scene new user needs and requirements related to the use of large amounts and varieties of data in different space domains such as Earth Observation (e.g., EU Copernicus programme), Space Science, Navigation and Telecommunications (e.g., EU Space programmes as Galileo and EU GovSatCom), mission operations and system engineering;
- Bring together major European actors in the fields of Space and data technologies, including research, industry, institutions, and users, to strengthen the communication and transfer of requirements, methods and technologies, and to reinforce an interdisciplinary approach;
- Explore and expand the ever increasing relevance of Big Data in European and global environmental policy initiatives and programmes, and the corresponding increasing complexity of applications and use cases;

- Discover and foster breakthrough data science processing and analysis techniques to extract insights and generate foresight, showing use cases wherever possible to facilitate future user uptake;
- Focus on new paradigms of data science addressing the entire value chain, i.e., building of reference training sets, data processing to extract information, information analysis to gather knowledge, and knowledge transformation in foresight;
- Maximise the uptake and impact of solutions exploiting multi-source spatio-temporal data linked with other data sources;
- Advance the upscale of new solutions from Research and Innovation (R&I) to operational use (e.g., for the security domain and informed policy making);
- Foster interoperability of platforms and services by promoting open standards, analysis ready data, and Application Programming Interfaces (APIs);
- Promote interdisciplinarity to respond to multi-sectorial challenges such as those put forward by the European Green Deal or the wide-ranging consequences of the Covid-19 pandemic;
- Promote cross-fertilisation with similar activities in other data intensive domains (e.g., high-energy physics, genomics, social media, internet of things, etc.).

The BiDS conference series is co-organised by the European Space Agency (ESA), the Joint Research Centre (JRC) of the European Commission, and the European Union Satellite Centre (SatCen). BiDS'21 emphasises not only on the insights that can be retrieved from Big Data from Space but also on the exploitation of these insights for foresight to improve our capacity to detect trends and model future evolution. This capacity is becoming increasingly important given the pace at which our World is changing. This is exemplified and reflected by the EU Destination Earth (DestinE) initiative and the related digital twin of the Earth. The objective of DestinE is to develop a very high precision digital model of the Earth to monitor and simulate natural and human activity, and to develop and test scenarios that would enable more sustainable development and support European environmental policies. The provision of more reliable scenarios of future evolution under different boundary conditions requires us to improve our understanding of Earth's dynamic systems besides their monitoring. Similarly to past editions of this conference, the 2021 edition provides a snapshot of the different research and innovation developments in the field of Big Data from Space including technical aspects and applications.

These proceedings contain the papers presented at the on-line BiDS'21 conference held on May 18-20 as an on-line conference. From a total of 63 submissions, reviewed in average by 3 programme committee members, 47 papers were accepted: 30 as oral presentations and the remaining 17 as poster presentations. Further to these presentations, the conference featured 5 keynote lectures from distinguished speakers that enlightened the audience with their experience in areas relevant to Big Data from Space:

1. *Digital Twins of the Earth System = Really Big Data*
   by Peter Bauer
   (European Centre for Medium-Range Weather Forecasts, UK)
2. *From Interactive Computing to Collaborative Science: Opportunities in the Cloud with Open Infrastructure*
   by Fernando Pérez
   (Berkeley Institute for Data Science, USA)
3. *Earth Observation + Machine Learning + System Modelling to Understand the Earth System*
   by Markus Reichstein (Max Planck Institute for Biogeochemistry, Germany)
4. *Space: the Quantum Frontier*
   by Radu Ionicioiu
   (Horia Hulubei National Institute for R&D in Physics and Nuclear Engineering, Romania)
5. *Big Data Astronomy: Challenges and Opportunities*
   by Leanne Guy
   (AURA/Rubin Observatory, USA)

BiDS'21 was initially scheduled to be hosted at the Polytechnic University of Bucharest, Romania, with the support of the Romanian Space Agency (ROSA). This venue was announced during the closing session of BiDS'19 organised in November 2019. No one could have imagined the disruption that we would all have to face just 3 months later with the covid-19 pandemic. Accordingly, the BiDS'21 organising committee took the decision to move the conference to an on-line event during its first meeting in May 2020. This was not an easy decision given the enthusiasm from our colleagues from Romania to host the conference and promote Big Data from Space in Romania, particularly among students and young researchers. Retrospectively, this was the right decision as most of us are still limited in travelling not to speak of gathering at a conference that attracted 700 attendees in 2019! Each conference format has its own advantages and drawbacks. The creation of new contacts and sparkling new ideas and collaborations during coffee break, lunches, and dinners have for sure been very much missed. On the plus side, all on-line presentations including the recording of the keynote lectures not included in these proceedings can be accessed through the conference website: www.bigdatafromspace2021.org.

Great thanks go to all authors and presenters of BiDS'21 as well as the numerous participants (518 registrations from more than 54 different countries one week ahead of the conference start). Together, they have ensured the success of the 2021 conference on Big Data from Space. Special thanks go to the Programme Committee members who have guaranteed the quality of the conference programme and these proceedings.

Pierre Soille, Sveinung Loekken, and Sergio Albani

Conference Chairs

Sergio Albani	European Union Satellite Centre (SatCen)
Sveinung Loekken	European Space Agency (ESA)
Pierre Soille	European Commission, Joint Research Centre (JRC)

Organising Committee

Sergio Albani	European Union Satellite Centre (SatCen)
Simon Baillarin	Centre National d'Études Spatiales (CNES), France
Mihai Datcu	German Aerospace Center (DLR), Germany
Jean-Pierre Gleyzes	Centre National d'Études Spatiales (CNES), France
Jutta Graf	German Aerospace Center (DLR), Germany
Pieter Kempeneers	European Commission, Joint Research Centre (JRC)
Sveinung Loekken	European Space Agency (ESA)
Vicente Navarro	European Space Agency (ESA)
Antonio Romeo	RHEA Group supporting ESA
Francesco Sgamarella	European Space Agency (ESA)
Pierre Soille	European Commission, Joint Research Centre (JRC)
Juan Luis Valero	European Union Satellite Centre (SatCen)
Joost van Bemmelen	European Space Agency (ESA)
Raffaelle Vitulli	European Space Agency (ESA)

Programme Committee

Selim Aksoy	Bilkent University, Turkey
Mirko Albani	European Space Agency (ESA)
Sergio Albani	European Union Satellite Centre (SatCen)
Christophe Arviset	European Space Agency (ESA)
Simon Baillarin	Centre National d'Études Spatiales (CNES), France
Francesco Barbato	European Commission, DEFIS
Peter Baumann	Jacobs University Bremen, Germany
Rogerio Bonifacio	World Food Programme
Francesca Bovolo	Fondazione Bruno Kessler, Italy
Lorenzo Bruzzone	University of Trento, Italy
Francesco Casu	IREA, National Research Council (CNR), Italy
Marco Chini	Luxembourg Institute of Science and Technology
Massimo Ciscato	European Commission, HaDEA
Esther Conway	Science and Technology Facilities Council, UK
Christina Corbane	European Commission, Joint Research Centre (JRC)
Raphaël d'Andrimont	European Commission, Joint Research Centre (JRC)
Mihai Datcu	German Aerospace Center (DLR), Germany
Begum Demir	TU Berlin
Jean Dusart	European Commission, DG Research and Innovation

Saso Dzeroski	Jozef Stefan Institute, Slovenia
Liang Feng	Universuty of Edinburgh, UK
Paolo Gamba	University of Pavia, Italy
Ana Garcia Robles	Big Data Value Association (BDVA)
Jean-Pierre Gleyzes	Centre National d'Études Spatiales (CNES), France
Jutta Graf	German Aerospace Center (DLR), Germany
Jacopo Grazzini	European Commission, DIGIT
Harm Greidanus	European Commission, Joint Research Centre (JRC)
Christophe Honvault	European Space Agency (ESA)
Vangelis Karkaletsis	NCSR "Demokritos", Greece
Pieter Kempeneers	European Commission, Joint Research Centre (JRC)
Panagiotis Kikiras	European Defence Agency (EDA)
Doris Klein	German Aerospace Center (DLR), Germany
Manolis Koubarakis	National and Kapodistrian University of Athens
Riccardo Lanari	IREA, National Research Council (CNR), Italy
Michele Lazzarini	European Union Satellite Centre (SatCen)
Bertrand Le Saux	European Space Agency (ESA)
Sébastien Lefèvre	Université de Bretagne Sud, France
Sveinung Loekken	European Space Agency (ESA)
Adrian Luna	European Union Satellite Centre (SatCen)
Luis Mansilla	European Space Agency (ESA)
Michele Manunta	IREA, National Research Council (CNR), Italy
Ana Martinez	European Commission, Joint Research Centre (JRC)
Gema Maza	European Union Satellite Centre (SatCen)
Lewis Mcgibbney	National Aeronautics and Space Administration (NASA)
Sabri Mekaoui	European Commission, HaDEA
Katrin Molch	German Aerospace Center (DLR), Germany
Vicente Navarro	European Space Agency (ESA)
Axel-Cyrille Ngonga Ngomo	Paderborn University, Germany
Allan A. Nielsen	Technical University of Denmark
Simon Oliver	Geoscience Australia
Edzer Pebesma	Inst for geoinformatics, University of Muenster, Germany
Jean-François Pekel	European Commission, Joint Research Centre (JRC)
Thierry Ranchin	MINES ParisTech
Antonio Romeo	RHEA Group, Italy
Rui Santos	European Space Agency (ESA)
Darek Saunders	European Border and Coast Guard Agency (FRONTEX)
Flavio Sbardellati	European Global Navigation Satellite Systems Agency (GSA)
Michael Schick	EUMETSAT
John Schnase	NASA
Francesco Sgaramella	European Space Agency (ESA)
Pierre Soille	European Commission, Joint Research Centre (JRC)
Peter Strobl	European Commission, Joint Research Centre (JRC)
Vasileios Syrris	European Commission, Joint Research Centre (JRC)
Olaf Trieschmann	European Maritime Safety Agency (EMSA)
Devis Tuia	Ecole polytechnique Fédérale de Lausanne (EPFL)
Juan Luis Valero	European Union Satellite Centre (SatCen)
Joost van Bemmelen	European Space Agency (ESA)

Stijn Vermoote	European Centre for Medium-Range Weather Forecasts (ECMWF)
Raffaele Vitulli	European Space Agency (ESA)
Julia Wagemann	European Centre for Medium-Range Weather Forecasts (ECMWF)
Wolfgang Wagner	Vienna University of Technology
Gui-Song Xia	Wuhan University, China
Xiaoxiang Zhu	German Aerospace Center (DLR), Germany

Additional Reviewers

Omar Barrilero	European Union Satellite Centre (SatCen)
Ridvan Kuzu	Technical University of Munich, Germany
Ivica Obadic	Technical University of Munich, Germany

Advisory Committee

Daniela Faur	Polytechnical University of Bucharest, Romania
Andreea Griparis	Polytechnical University of Bucharest, Romania
Irina Manciu	Romanian Space Agency (ROSA)
Corina Vaduva	Polytechnical University of Bucharest, Romania

Table of Contents

DEEPCUBE: EXPLAINABLE AI PIPELINES FOR BIG COPERNICUS DATA

Ioannis Papoutsis[1], Alkyoni Baglatzi[1], Souzana Touloumtzi[1], Markus Reichstein[2], Nuno Carvalhais[2], Fabian Gans[2], Gustau Camps-Valls[3], Maria Piles[3], Theofilos Kakantousis[4], Jim Dowling[4], Manolis Koubarakis[5], Dimitris Bilidas[5], Despina-Athanasia Pantazi[5], George Stamoulis[5],Christophe Demange[6], Léo-Gad Journel[6], Marco Bianchi[7], Chiara Gervasi[7], Alessio Rucci[7], Ioannis Tsampoulatidis[8], Eleni Kamateri[8], Tarek Habib[9], Alejandro Díaz Bolívar[9], Zisoula Ntasiou[10], Anastasios Paschalis[10]

[1]National Observatory of Athens, Institute for Astronomy, Astrophysics, Space Applications & Remote Sensing, [2]Max Planck Institute for Biogeochemistry, [3]University of Valencia, Image Processing Laboratory, [4]Logical Clocks AB, [5]National and Kapodistrian University of Athens, Department of Informatics and Telecommunications, [6]GAEL Systems, [7]TRE ALTAMIRA, CLS Group, [8]INFALIA [9]MURMURATION SAS - Flockeo.com, [10]Hellenic Fire Service

ABSTRACT

The H2020 DeepCube project leverages advances in the fields of Artificial Intelligence and Semantic Web to unlock the potential of Copernicus Big Data and contribute to the Digital Twin Earth initiative. DeepCube aims to address problems of high socio-environmental impact and enhance our understanding of Earth's processes correlated with Climate Change. To achieve this, the project employs novel technologies, such as the Earth System Data Cube, the Semantic Cube, the Hopsworks platform for distributed deep learning, and visual analytics tools, integrating them into an open, cloud-interoperable platform. DeepCube will develop Deep Learning architectures that extend to non-conventional data, apply hybrid modeling for data-driven AI models that respect physical laws, and open up the Deep Learning black box with Explainable Artificial Intelligence and Causality.

Index Terms— Data cubes, Artificial Intelligence, semantic web, hybrid modeling, explainable AI, causality, climate change, Digital Twin Earth

1. INTRODUCTION

The Copernicus program is believed to be a game changer for both science and the industry. Free and open data available at this scale, frequency, and quality constitutes a fundamental paradigm change in Earth Observation (EO). However, the availability of the sheer volume of Copernicus data outstrips our capacity to extract meaningful information. The EO community needs technology enablers to propel the development of entirely new applications at scale.

Deep Learning (DL) has been one of the fastest-growing trends in big data analysis. It is only relatively recently that DL was introduced to the EO research community for information extraction from big satellite data. The majority of the applications that use DL though, seem to reiterate old EO problems, which now can be solved faster and provide incrementally higher accuracy with respect to conventional Machine Learning (ML) approaches.

Furthermore, DL leads to highly nonlinear, overparameterized models. They excel in prediction accuracy, but such complexity hampers interpretability and trustworthiness. Predictive accuracy is important but often insufficient, and interpreting what the models learned becomes important, especially in problems with economical, societal or environmental implications. The lack of interpretability, i.e. the degree to which a human can understand the cause of a decision has become a main barrier of DL in its wide-spread applications for geosciences.

Finally, EO data becomes useful only when analyzed together with other sources (e.g., geospatial & in-situ data) and turned into knowledge. Linked data is a data paradigm that studies how one can make Resource Description Framework (RDF) data available on the web and interconnect it with other data with the aim of increasing its value. Nevertheless, there are only a handful of applications that showcase the semantic integration of linked EO and non-EO products.

The H2020 DeepCube project (Jan. 2021 - Dec. 2023, https://deepcube-h2020.eu/) leverages advancements in the fields of AI and semantic web to unlock the potential of big Copernicus data. It aims to address problems that imply high environmental and societal impact, enhance our understanding of Earth's processes, correlated with the climate emergency, and feasibly generate high business value, in line with the Destination Earth and the Digital Twin Earth objectives. To achieve this, DeepCube integrates mature and new technologies into an open interoperable platform that can be deployed in cloud environments, DIAS included. The platform is then used to develop novel DL pipelines to extract value from big Copernicus data. DeepCube develops

doi:10.2760/125905

DL architectures that extend to non-conventional data and problems, introduces a novel hybrid modeling paradigm for data-driven AI models that respect physical laws [1], and opens-up the DL black box through Explainable AI (XAI) and Causality. We showcase these in six applications.

2. TECHNOLOGIES

DeepCube makes use of mature technology enablers that have been developed in other European Commission and European Space Agency funded research. In DeepCube these enablers are integrated to an interoperable environment allowing EO and AI specialists to create value chains from a wide offer of raw EO and non-EO big data. This environment is the DeepCube platform (Fig. 1), which will scale to big Copernicus datasets, designed to share resources and to define dataflows in a coherent integrated solution. DeepCube platform will be deployed into more than one cloud environments, including Copernicus DIAS. Its individual components are briefly described next.

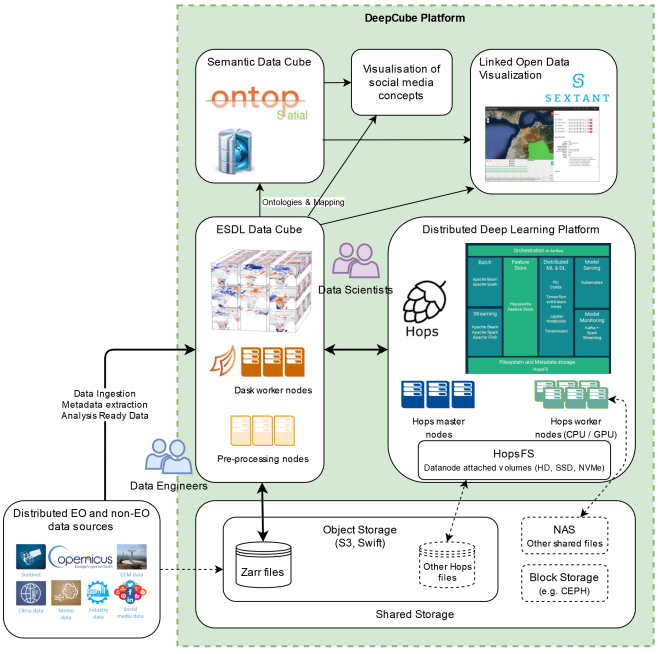

Fig. 1. High level architecture of the DeepCube platform.

The **Earth System Data Cube (ESDC)** developed by Earth System Data Lab project, seeks to be a service to the scientific community to facilitate access and exploitation of multivariate data sets in Earth Sciences to actually understand the interactions between the Earth's subsystems.The core part of the ESDC is the data in analysis-ready form, together with tools and methods to generate, access, and exploit the ESDC. A data cube essentially consists of screened, or Analysis Ready Data (ARD), with the dimensions "latitude", "longitude", "time", "variable". Further dimensions can be added as a result of an analysis. Currently ESDC supports a common spatio-temporal grid [2], DeepCube will advance to create Data Cubes where information layers are stored in heterogeneous spatio-temporal resolution. ESDC is committed to open source computations, and open data usage. Dynamic resource allocation and rapid scalability of ESDC are its cornerstones for data analysis on the cloud.

The **Semantic EO Data Cube** [3] enables the semantic enrichment of ESDC. The Semantic Data Cube allows users to query metadata, EO data, other Linked Open Data (LOD), and information/knowledge extracted from the data using a semantic query language, thus creating new value chains. In a semantic data cube [3], at least one categorical interpretation exists for each observation in an image (i.e., each pixel). EO data co-exists with its interpretation and can also be queried using the same high-level query language that is used for querying interpretations. For example, a user can query the reflectance values of certain bands in an image (e.g., for calculating an index) and in the same query also refer to an interpretation of these values. Semantic data cubes are also enriched by other kinds of data (e.g., other kinds of geospatial data such as OSM). In this way, the querying possibilities for a user become even larger. DeepCube will develop the first semantic data cube technology internationally by extending the geospatial ontology-based data access system Ontop-spatial [4].

Hopsworks is an open-source **Data-Intensive AI** platform for developing and operating end-to-end ML pipelines at scale. Hopsworks provides first-class support for popular open-source frameworks for distributed data processing, data engineering and data science. In addition, Hopsworks supports DL on large volumes of data, such as those produced by the Copernicus program, using distributed training. Distributed training uses many GPUs and data-parallel model training to reduce the time required to train models by adding more GPUs. Hopsworks leverages Apache Spark to make distributed training easier for programmers. However, modern approaches to distributed training require developers to rewrite their code when moving from using a single GPU to hyperparameter tuning (using lots of GPUs) to distributed training. DeepCube will develop a comprehensive new framework that unifies single-host training, hyperparameter tuning, and distributed training. We will also expand Hopsworks to support model-parallel training, as well as API support for distributed semi-supervised learning and self-supervised learning. As such, DeepCube will build a state-of-the-art and the most feature complete framework for distributed DL.

DeepCube will extend **Sextant** [5], a web based and mobile ready platform for **visualizing**, exploring and interacting with linked geospatial data. Sextant is a user-friendly application that allows both domain experts and non-experts to take advantage of semantic web technologies, creating thematic maps by combining spatio-temporal information with other data sources, e.g. industrial intelligence, socio-economic

doi:10.2760/125905

data, etc., allowing visual analytics based on big Copernicus data. In addition, DeepCube will develop user interfaces offering multiple ways of visualisation and filtering of social media data, detected locations and visual concepts, allowing analytics on top of them.

3. APPLICATIONS

3.1. Forecasting localised drought impacts in Africa

Climate change will lead to an accumulation and intensification of various climate extremes [6]. Drought and heat waves, as experienced repeatedly in the last decade, are expected to become more frequent in the future, as the corresponding persistent weather situations become more and more probable. The effects on various sectors are substantial, as could be seen, for example, from the effects on agriculture, inland waterways, and consequently nutrition and energy supply.

There are two significant gaps that will be addressed by DeepCube: the first one relates to lack of methods for assessing, in fine resolution, drought impact at the local level. This requires downscaling from meteorological scales to sub-km level using satellite data. The second gap is a lack of understanding of memory effects considering ecosystem dynamics, after a drought event. A better understanding will be achieved with so-called hybrid dynamic models [1], which model the system partly with physical equations, partly with ML.

3.2. Climate induced migration in Africa

In the current context of climate change, extreme heat waves, droughts and floods are not only impacting the biosphere and atmosphere but the anthroposphere too. Human populations are forcibly displaced, which are now referred to as climate-induced migrants. On the agenda of the United Nations Framework Convention on Climate Change, for instance, there is an item dedicated to migration, displacement and human mobility. The problem has obvious environmental, societal and economic implications, in both adaptation and mitigation to climate change, as well as for assistance to their home states. Modeling, anticipating, characterizing and understanding the severity of migration flows and the direct and latent factors are of paramount relevance.

There is a growing number of media reports assuming the link of climate change, conflicts, and forced migration. However, there is little empirical evidence supporting that climate change and migration are interrelated [7]. At present, there is no theoretical approach to adequately represent the causal mechanisms through which climate change induces human displacement and migration flows. This will be the first time that advanced causal inference schemes are developed to investigate the climate-induced migration in Africa.

Therefore, DeepCube will identify the main environmental and socio-economic drivers of human mobility and develop models able to reproduce and forecast migration flows, apply causal discovery methods to gain a deeper understanding of the characteristics of the climate-induced migration flows and establish the causal relationships of environmental and socio-economic drivers with human mobility in sub-Saharan Africa.

3.3. Fire hazard forecasting in the Mediterranean

Climate change is playing an increasing role in determining wildfire regimes, with future climate variability expected to enhance the risk and severity of wildfires in many biomes including Southern Europe [6]. Fire hazard forecasting systems linked with the operational authorities (Civil Protection, Fire Brigade/Service etc.), would increase their preparedness and enhance the emergency response capacity in a changing climate.

DeepCube will identify the climatic, vegetation status and anthropogenic drivers that impact the most fire proneness based on multivariate historical data analysis on the Mediterranean. Based on these insights, the application will use AI bound by an ecosystem modeling [1] to model short and mid-term fire hazard and make more accurate and with less uncertainty future predictions using EO data time-series analysis. XAI techniques (permutation analysis, visualization of features-heatmap activations, and clustering activations) will be used to open-up the DL box and gain trust on what the model has learnt. Finally fire hazard forecasts will be combined with LOD to assess fire risk for assets (population, environment, economic activity) on the ground.

3.4. Global volcanic unrest detection & alerting

Interferometric Synthetic Aperture Radar (InSAR) can systematically provide ground deformation estimations over volcanic areas, see 6-day repeat pass cycle of Sentinel-1AB. Fringes detected in Sentinel-1 wrapped interferograms over volcanic areas indicate the onset of deformation, which is usually due to magma chamber fill-in at depth. Such activity is considered as precursor for a potential eruption.

Having the work by Anantrasirichai et al. [8], as a starting point, DeepCube will research DL architectures that can automatically detect the presence of ground deformation triggered by volcanic unrest, within wrapped interferograms, towards establishing a volcanic deformation alert service, covering several volcanoes globally.

3.5. Automated infrastructure monitoring with InSAR

SAR-derived information is used to produce millimetric-precision ground surface deformation maps. Thanks to Sentinel-1 SAR revisit time, new deformation maps can be delivered to end-users on a regular basis [9], showing average deformation rates (mm/yr) of Persistent Scatter (PS) “points” and their displacement time series. Each information layer is made of hundreds of thousands of measurement points,

doi:10.2760/125905

and can be used for detecting significant instabilities on critical infrastructures thus contributing to plan and optimize mitigation actions.

However, no automated processes are in place to robustly detect hotspots, i.e. zones for which displacement time series show a significant change in trend motion. In addition, for zones experiencing these changes, no indication is given to end-users about possible reasons and driving mechanisms. DeepCube will attempt to link any deformation hotspots to a possible reason for trend change, using DL on InSAR data and sparse in-situ geodetic measurements for training and fusion.

3.6. Copernicus services for sustainable tourism

Tourism is one of the pillars of the modern economy. It constitutes more than 10% of global GDP with a CAGR of 3+%. The number of international tourists is forecasted to rise to 1.8 Billion in 2030, making it crucial to find efficient ways to handle this growth, preserve the fragile destinations and adapt to the increasing demand over limited hospitality infrastructures. Additionally, more than 65% of European travellers have declared that they are striving to make their travels more sustainable but do not find the right information or the possibility to assess their environmental footprint.

DeepCube will create a new commercial service, by producing a pricing engine for tourism packages, which incorporates the environmental dimension. The goal is to calculate a suite of price coefficients for a travel agency to apply to its packages, considering environmental impact automatically, utilizing Copernicus and data (water quality degradation, marine pollution, air pollution), product characteristics (ecological potential), and supply and demand information coming from social media streams. The application will be set-up as a reinforcement learning problem and a prototype will be developed for the northeast coast of Brazil nearby the Lencois national park.

4. CONCLUSIONS

We see DeepCube as a showcase of the Digital Twin Earth potential, by 1) delivering the DeepCube platform as a technology enabler for the deployment of end-to-end AI pipelines on big EO data regardless of the underlying cloud infrastructure, and 2) designing and testing new AI architectures to address significant scientific questions related to Climate Change and generating business value via the joint analysis of EO with industrial data.

The DeepCube platform consists of mature, high technology readiness level, components. This interoperable platform will be a DeepCube legacy which could be deployed in different cloud environments. The platform will provide novel solutions for all phases on an EO-based AI pipeline, from data ingestion, to big data organisation (data cubes), feature engineering, semantic annotation, distributed DL, semantic reasoning and visualisation.

In addition, DeepCube will test a hybrid modeling approach for geophysical parameters estimation, enhanced through XAI for “physics-aware” AI applications. DeepCube will also perform causality analysis to understand and interpret patterns, cause and effects on diverse datasets, including satellite, social media and socio-economic data. Finally, it will employ for the first time AI on complex Sentinel-1 SAR data, an archive of the order of PBs, currently the richest asset that remains underexploited. We expect that the first concrete results will be shared by the end of 2021.

DeepCube will deliver to the community Data Cubes with ARD and training datasets allowing to capture hidden trends for key environmental variables. These cubes will be made available for reuse by June 2021.

REFERENCES

[1] Reichstein, M., Camps-Valls, G., Stevens, B. et al. ”Deep learning and process understanding for data-driven Earth system science”, Nature, 566, 195–204, doi: 10.1038/s41586-019-0912-1, 2019

[2] Mahecha, M. D., Gans, F., Brandt, G., et al. ”Earth system data cubes unravel global multivariate dynamics”, Earth Syst. Dynam., 11, 201–234, doi: 10.5194/esd-11-201-2020, 2020.

[3] Augustin, H., Sudmanns, M., Tiede, D., Lang, S., Baraldi, A. ”Semantic Earth Observation Data Cubes”, Data, 4, 102. doi: 10.3390/data4030102, 2019.

[4] Bereta K., Xiao G., Koubarakis M. ”Ontop-spatial: Ontop of geospatial databases”, Journal of Web Semantics, 58, doi: 10.1016/j.websem.2019.100514, 2019.

[5] Nikolaou C., Dogani K., Bereta K., Garbis G., Karpathiotakis M., Kyzirakos K., Koubarakis M., ”Sextant: Visualizing time-evolving linked geospatial data”, Journal of Web Semantics, 35, 1, 35-52, doi: 10.1016/j.websem.2015.09.004, 2015

[6] Shukla P.R., Skea J., Calvo Buendia E., et al. ”IPCC, 2019: Climate Change and Land: an IPCC special report on climate change, desertification, land degradation, sustainable land management, food security, and greenhouse gas fluxes in terrestrial ecosystems”, 2019.

[7] Brzoska M., Fröhlich C., ”Climate change, migration and violent conflict: vulnerabilities, pathways and adaptation strategies, Migration and Development”, 5:2, 190-210, doi: 10.1080/21632324.2015.1022973, 2016.

[8] Anantrasirichai N., Biggs J., Albino F., Bull D., ”A deep learning approach to detecting volcano deformation from satellite imagery using synthetic datasets”, Remote Sensing of Environment, 230, doi: 10.1016/j.rse.2019.04.032, 2019.

[9] Raspini F., Bianchini S., Ciampalini A., et al., ”Continuous, semi-automatic monitoring of ground deformation using Sentinel-1 satellites” Scientific Reports, 8:7253, doi: 10.1038/s41598-018-25369-w, 2018.

CORRELATION BETWEEN SATELLITE AND IN-SITU MEASUREMENTS, STUDYING CO AND CO2 OBSERVATIONS FROM SENTINEL-5P, OCO2 AND ICOS IN-SITU DATA

Alejandro Diaz & Tarek Habib

Murmuration - 15 rue Victor Hugo, 31150 Bruguières

ABSTRACT

Continuous monitoring of Greenhouse Gases (GHGs) is of vital importance as it is key to measure our progress towards the achievement of the sustainability of our environment. The same applies to global atmospheric simulation of GHGs variations and fluxes that enables us to understand and build the future climatic scenarios. Both tasks are complex requiring high computational costs, human validation work and a large number of geographically distributed measurement instruments. ICOS, the European Integrated Carbon Observation System is a distributed international research infrastructure dedicated to measure, analyze and understand those fluxes. On the other hand, the space agencies ESA and NASA have developed in the last decade satellites capable of measuring air quality, through the observation of GHGs, with increased accuracy. Satellites are capable of measuring the concentration of GHGs in the troposphere. This paper presents our findings in computing the correlation between ground-based data using the ICOS infrastructure data with OCO2 and S5P satellite data to validate the use of satellite to improve our global observation capacity and the global simulation models of GHGs. The paper shows the potential of using space based observations to characterize sources and sinks of GHGs on a local scale.

Index Terms— Copernicus, in-situ sensors, Sentinel-5P, OCO2, ICOS, CO, CO2, air quality, green house gases

1. INTRODUCTION

Over the past three decades, several international treaties, policies and collaboration frameworks have been put in place to monitor, regulate and reduce the global Greenhouse Gases (GHGs) emissions. UNFCCC (United Nations Framework Convention on Climate Change), the Kyoto protocol [3] and the Paris agreements [4] are few examples of this international consensus regarding the importance of reducing the GHGs on a global scale. The efficient monitoring and modeling procedures needs efficient, reliable and dense observations. This is a global effort where each country, region, or coalition, work to put in place an observation network with a variety of sensors. To converge on a common understanding between all stakeholders, these networks have to be inter-operable and the sensors compatible. Another possibility is to put in place global monitoring sensors, like earth observation satellites. The challenge is to validate the use of these global observations with local instruments to be able to identify the sources and sinks of GHGs at the finest scale, with the highest reliability and at the highest revisit, temporal frequency.

The Copernicus program is the most ambitious environment monitoring program to date. With its variety of satellite sensors and in-situ components, the program provides unprecedented monitoring at a global scale. Regarding GHGs, the Sentinel-5P (S5P) satellite carries the state-of-the-art Tropomi instrument to map a multitude of gases including CO. In addition to S5P, the Orbiting Carbon Observatory, OCO-2, is NASA's first dedicated Earth remote sensing satellite to study atmospheric carbon dioxide (CO2) from space. Despite their differences in geographical resolution and revisit frequency, the combination of data coming from these two satellites, S5P and OCO-2, provides a powerful global monitoring tool of GHGs.

While satellite data are key to understand the global dynamics; the local fluxes, the sectorial contributions and the national inventories require finer geographical resolutions. The Integrated Carbon Observation System (ICOS) is a research infrastructure aiming to quantify the GHGs balance of Europe and adjacent regions. Through the deployed ground stations, ICOS, brings in reliable, scientific data to help decision making bodies reach the consensus related to GHGs and efficiently allocate the required efforts to reduce them.

The use of satellite data for global understanding and local in-situ sensors for local observations is proven to be very relevant to support decision making, nonetheless, the combination of both technique to increase our monitoring and modelling capacities is a high priority research area. Scientific effort have been put lately in linked subjects, for instance, for the validation of NO2 S5P data with ground data from NDACC ZSL-DOAS, MAX-DOAS and Pandonia global networks [6] [5]. The results show clear correlations between the NO2 satellite and in-situ data. Another example is the use of S5P data to monitor GHGs in cities have been studied and proven reliable in [7]. Finally the cross validation of satellite data and portable in-situ spectrometer have been explored and demonstrated in [8]. All these recent elements give us confi-

doi:10.2760/125905

dence in the relevance of this study and the need to explore the possibility to use satellite measurements for local inventory processes.

The challenge we are addressing in this paper is to assess the capacity for satellite measurement techniques, in the troposphere, to correlate to local, in-situ, sources and sinks. We are presenting our work to study the correlation between satellite measurements from S5P, OCO-2 and ICOS ground stations, opening the way towards potential use of satellite measurements as a source of virtual densification of the in-situ measurements networks.

Following this introduction, the paper is divided in 4 chapters, starting with the presentation of the data sources and the datasets we have used, the Data analysis and the associated processing pipeline, the results and finally the conclusions and future work.

2. DATA SOURCES

2.1. ICOS

ICOS is a research infrastructure that has been born out of European scientific communities' idea of having a consistent, sustained measurement network operating under exactly the same technical and scientific standards to enable high-quality climate change research and increase usability of the research data.

The ICOS Atmosphere network includes stations in 13 European countries. Each station measures greenhouse gas concentrations (such as carbon dioxide and methane) in the atmosphere and is standardised by using the same equipment, technology and a rigorous quality control process [9]. The data collected at the atmosphere stations are automatically processed and quality controlled by the Atmoshpere Thematic Centre, checked by the station Principle Investigator and finally published via the ICOS Carbon Portal.

In this article we have made use of data published on the ICOS Carbon portal from 25 ICOS approved Class 2 stations through its API [14], using the CO and CO2 concentration values at the highest sampling level for each of them. The image Fig. 1 shows the location in Europe of the 25 stations used.

Fig. 1. Location in Europe of the 25 Class 2 ICOS stations

2.2. Sentinel-5P

Sentinel-5 Precursor (S5P) is an Earth observation satellite developed by ESA as part of the Copernicus Programme and launched on 13 October 2017 to monitor air pollution. Through its Tropomi spectrometer (TROPOspeheric Monitoring Instrument) the satellite monitors ozone, methane, formaldehyde, aerosol, NO2, SO2 and the gas of interest for this article, CO. To observe this molecule at a global level, TROPOMI exploites clear-sky and cloudy-sky Earth radiance measurements in the 2.3 μm spectral range of the shortwave infrared (SWIR) part of the solar spectrum. TROPOMI clear sky observations provide CO total columns with sensitivity to the tropospheric boundary layer. For cloudy atmospheres, the column sensitivity changes according to the light path [10].

The original S5P Level 2 (L2) data is binned by time, not by latitude/longitude. To facilitate the analysis, the Earth Engine Data Catalog, developed by Google (GEE), has been used in this paper as a source of data, converting each L2 product into an L3 but maintaining a single grid per orbit [12].

2.3. OCO2

The Orbiting Carbon Observatory, OCO2, is the NASA's first dedicated Earth remote sensing satellite to study atmospheric carbon dioxide from Space. OCO2 is collecting space-based global measurements of atmospheric CO2 with the precision, resolution, and coverage needed to characterize sources and sinks on regional scales. OCO2 is also able to quantify CO2 variability over the seasonal cycles year after year.

The JPL presents in the Algorithm Theoretical Basis (ATB) [11] the algorithm to compute the column-averaged CO2 dry air molecule fraction X_{CO2}, the variable chosen in this paper to be correlated with the in-situ data from the ICOS infrastructure. In addition, from the CO2 Virtual Science Data Environment [13], a NASA web service, JPL allows to query level 2 or 3 products, the second ones derived from downsampling the original data at the desired spatial and temporal resolution.

3. DATA ANALYSIS ENVIRONMENT

To perform the temporal correlation at each of the ICOS stations, it is necessary to extract the corresponding data from each satellite and from the ICOS infrastructure itself. As GEE and ICOS provide a python API, we have setup a python environment for the post-processing of these data and the calculation of the correlation between them. Fetching OCO2 dataset has been performed using independent scripts and the data uploaded to the Google platform, making the corresponding transformations to make the data type compatible.

Regarding data extraction, the calculation has been performed for each of the class 2 stations of the ICOS infrastructure. Therefore we filtered the level 3 products of OCO2 and S5P for the position given by each station by taking the start

doi:10.2760/125905

and end dates of the query as the limit dates of operation of the station, which generally range from the validation of the class 2 label of each tower to the present.

Using a python script and the previously mentioned ICOS and GEE APIs, time series have been extracted for each station for both, CO and CO2. To be able to use the GEE API and use the CO2 image collection from OCO2, a transformation of the level 3 product obtained from the NASA web platform in Netcdf4 format (generated by the NASA server as a derivative of the product "OCO2 Level 2 Full Physics Retrieval" [13]) has been carried out to obtain a series of TIF images that have been subsequently geo-referenced (geoTIF) and finally uploaded to the google server.

To setup the correlation, the temporal resolution of each data source is different, ICOS stations can provide hourly measurements, S5P a daily observation and OCO2 a monthly observation. After extracting the time series with the maximum possible sampling from each of the 3 data sources, these series have been resampled by an arithmetic mean. CO data have been resampled to a daily average to correlate ICOS and S5P data, while CO2 data have been resampled to a monthly mean to correlate ICOS and CO2 data. Thus, we have 4 time series for each ICOS station location, 2 for each gas (CO and CO2), both from satellites and ICOS.

A pairwise correlation between each time series performed for each station. We found erroneous and inconclusive data for locations where there are very few months of overlap between satellite and in-situ acquisitions. Moreover, the values at each location have been concatenated to compute an averaged correlation, filtering the series to only those with more than 1 year of coexistent satellite-in-situ data.

All the code necessary for reproducibility has been uploaded to Murmuration's Github [1].

4. RESULTS

The results in tables 1 and 2 show the correlation values between the satellite and in-situ time series, the number of months of coexistence of both sensors and the % of invalid values. The invalid satellite data are mainly due to the orbital nature of the satellite itself and the reduced frequency of visits for high latitudes (northern Europe) or to periods of inoperability of the in-situ station.

For the CO, the correlation values for the 25 stations range from low (0.55) to high (0.90) excluding outliers. Table 1 collects the top-3 and bottom-3 correlation values, as well as the average correlation of the 25 time series.

Fig. 2 and Fig. 3 show the daily evolution of CO for the station with the highest and lowest correlation value, respectively.

Table 2 shows the results for CO2. Considering the high number of invalid values in most of the stations, the number of valid rows is greatly reduced. In the table, the top-3 and the bottom-3 of the stations for which there is a minimum of 12 months of satellite-ICOS coexistence are listed. The absence of OCO2 data is high due to the low temporal resolution and the orbital nature of the satellite. Figure 4 shows the evolution of CO2 in the station with the highest level of correlation.

Table 1. Correlation of CO taking S5P and ICOS measurements. The column **Months** represents the amount of time where both sensors coexist. The % of **NaN** values from each of the data sources indicates the % of empty values within the coexistence interval of both, satellite and ICOS station.

	Correlation	**Months**	**%NaN SAT**	**%NaN ICOS**
PUY	0.8936	19	0	0
CMN	0.8662	19	0	0
RUN	0.8612	19	0	0
KRE	0.2926	19	0	15.8
LIN	0.2594	19	0	0
SAC	0.2371	19	0	0
Mean	0.6834	-	-	-

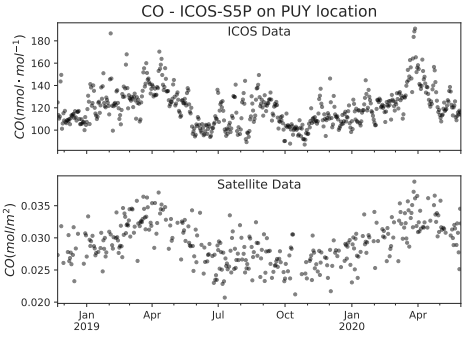

Fig. 2. Evolution of CO values measured at Puy-de-Dôme Observatory (PUY) in-situ and via S5P

5. CONCLUSIONS

Based on the two correlation tables and the study of the data through the associated graphs, it is concluded that the correlation is sufficiently high for both CO and CO2. It is important to note that although both the satellite and the in-situ sensor infer the same quantities, the satellite estimates the total amount of the molecule in the entire atmospheric column while the in-situ station accurately obtains the value of the molecule at a single point. Even so, the correlation between the two measurements is high, as has been shown throughout this article.

On the other hand, although the data have been well processed, both the ICOS infrastructure and the Copernicus project are young and the data history is not large. We have seen irrelevant correlations for time series of a few months

doi:10.2760/125905

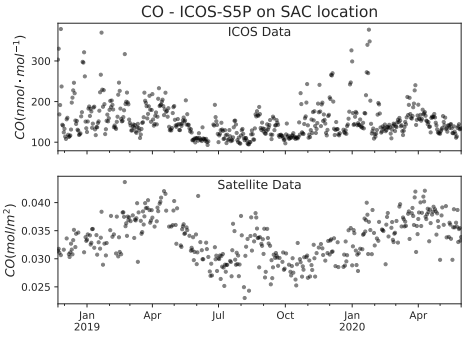

Fig. 3. Evolution of CO values measured at Saclay (SAC) in-situ and via S5P

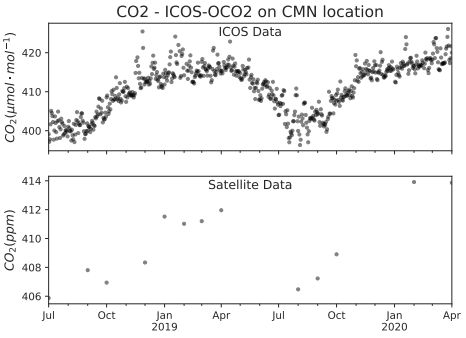

Fig. 4. Evolution of CO2 values measured at Monte Cimone Observatory (CMN) in-situ and via OCO2

Table 2. Correlation of CO2 taking OCO2 and ICOS measurements. **Months** and **NaN** are defined as for 1

	Correlation	**Months**	**%NaN SAT**	**%NaN ICOS**
CMN	0.9043	22	40.9	0
IPR	0.8386	27	37.0	0
LIN	0.8191	49	65.3	0
KIT	0.6757	40	35.0	0
KRE	0.6458	37	54.1	8.1
JFJ	0.4164	39	53.8	0
Mean	0.6456	-	-	-

of data and we have also been forced to resample to monthly values to be consistent with the OCO2 satellite data.

This first study shows that such correlations and cross time series can be used to validate global CO and CO2 models, expecting at each earth point correlation values of the same order as those found at each station. Future work will also include other GHGs for comparison, such as CH4, which is calculated at each of the Class 1 and 2 ICOS stations, as well as by the Sentinel 5P satellite.

REFERENCES

[1] Alejandro Diaz and Tarek Habib , “Correlation of satellite greenhouse gases data with European ICOS ground infrastructure - Github repository” Github Repository, 2021

[2] Laurent, O., ICOS Atmosphere Monitoring Station Assembly, & ICOS Atmosphere Thematic Centre (ATC). (2017). ICOS Atmospheric Station Specifications v1.3. ICOS ERIC. doi.org/10.18160/SDW6-BX90, 2017

[3] UNFCCC (1997) Kyoto Protocol to the United Nations Framework Convention on Climate Change adopted at COP3 in Kyoto, Japan, on 11 December 1997. UNFCCC Kyoto Protocol,1997

[4] Paris Agreement (adopted 2015-12-12 entered into force 2016-11-04) United Nations Treaty Collection, Chapter XXVII 7. d

[5] De Mazière, M. et al., ”The Network for the Detection of Atmospheric Composition Change (NDACC): history, status and perspectives”,Atmospheric Chemistry and Physics, Volume 18,2018,Number 7, pages 4935–4964

[6] Verhoelst, T. and Compernolle, S. and Pinardi et al., ”Ground-based validation of the Copernicus Sentinel-5P TROPOMI NO2 measurements with the NDACC ZSL-DOAS, MAX-DOAS and Pandonia global networks”, Atmospheric Measurement Techniques,Volume 14,2021,Number 1, Pages 481–510

[7] Hayoung Park, Sujong Jeong, Hoonyoung Park, Lev D. Labzovskii, Kevin W. Bowman, ”An assessment of emission characteristics of Northern Hemisphere cities using spaceborne observations of CO2, CO, and NO2”,Remote Sensing of Environment, Volume 254,2021,112246,ISSN 0034-4257, , 2021

[8] Humpage, Neil and Boesch, Hartmut and Okello, William and Dietrich, Florian and Chen, Jia and Lunt, Mark and Feng, Liang and Palmer, Paul, “Greenhouse gas column observations from a portable spectrometer in tropical Africa” ICOS Science Conference, 2020, doi: 10.13140/RG.2.2.14528.97280

[9] COSRI (2020):ICOS Atmosphere Station Specifications V2.0(editor:O.Laurent).ICOS ERIC.

[10] J. Landgraf, (2018), “ATB Document for Sentinel-5 Precursor: CO Total Column Retrieval” ATB-document

[11] Crisp D, Bosch H, Brown L, et al. 2012. OCO (Orbiting Carbon Observatory)-2 level 2 full physics retrieval algorithm theoretical basis. Technical Report. OCO D-65488, NASA Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, version 1.0

[12] Sentinel-5P NRTI CO: Near Real-Time Carbon Monoxide Earth Engine Data Catalog

[13] CO2 Virtual Science Data Environment - “Build your data product” co2.jpl.nasa.gov

[14] ICOS Carbon Portal Python Package. “ICOS CarbonPortal API” icos-carbon-portal.github.io

ARTIFICIAL INTELLIGENCE AND BIG DATA TECHNOLOGIES FOR COPERNICUS DATA: THE EXTREMEEARTH PROJECT

Manolis Koubarakis[1], George Stamoulis[1], Dimitris Bilidas[1], Theofilos Ioannidis[1], George Mandilaras[1], Despina-Athanasia Pantazi[1], George Papadakis[1], Vladimir Vlassov[2], Amir H. Payberah[2], Tianze Wang[2], Sina Sheikholeslami[2], Desta Haileselassie Hagos[2], Lorenzo Bruzzone[3], Claudia Paris[3], Giulio Weikmann[3], Daniele Marinelli[3], Torbjørn Eltoft[4], Andrea Marinoni[4], Thomas Kræmer[4], Salman Khaleghian[4], Habib Ullah[4], Antonis Troumpoukis[5], Nefeli Prokopaki Kostopoulou[5], Stasinos Konstantopoulos[5], Vangelis Karkaletsis[5], Jim Dowling[2, 6], Theofilos Kakantousis[6], Mihai Datcu[7], Wei Yao[7], Corneliu Octavian Dumitru[7], Florian Appel[8], Silke Migdall[8], Markus Muerth[8], Heike Bach[8], Nick Hughes[9], Alistair Everett[9], Åshild Kiærbech[9], Joakim Lillehaug Pedersen[9], David Arthurs[10], Andrew Fleming[11], Andreas Cziferszky[11]

[1] National and Kapodistrian University of Athens [2] KTH Royal Institute of Technology, Stockholm [3] University of Trento
[4] UiT The Arctic University of Norway [5] National Center for Scientific Research - Demokritos [6] Logical Clocks AB
[7] German Aerospace Center (DLR) [8] VISTA Remote Sensing in Geosciences GmbH [9] Norwegian Meteorological Institute
[10] Polar View [11] British Antarctic Survey

ABSTRACT

ExtremeEarth is a three-year H2020 ICT research and innovation project. Its main objective is to develop Artificial Intelligence and big data technologies that scale to the large volumes of big Copernicus data, information and knowledge, and apply these technologies in two of the European Space Agency (ESA) Thematic Exploitation Platforms (TEP): Food Security and Polar.

Index Terms— ExtremeEarth, Earth Observation, Linked Geospatial Data, Artificial Intelligence, Deep Learning, Copernicus, Food Security, Polar Regions

1. INTRODUCTION

Copernicus data is a paradigmatic case of big data giving rise to all relevant challenges, the so-called 5-Vs: volume, velocity, variety, veracity, and value, as it is documented in recent reports, such as the 2019 Copernicus Sentinel Data Access Report and the Copernicus Market Report of the same year. Copernicus data today is freely available not only through the Copernicus Open Access Hub but also through the five Data and Information Access Services (DIAS), where computing power is also available close to the data. Some related facilities of the Earth Observation (EO) ecosystem in Europe are the Thematic Exploitation Platforms (TEPs) of the European Space Agency (ESA), which enable user communities to collaborate using a virtual workspace where EO data, non-EO data, tools, and computing power are available. Today most of the TEPs run on a DIAS (e.g., the Food Security and Polar TEPs run on CREODIAS).

This work is supported by the ExtremeEarth project funded by European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 825258.

ExtremeEarth[1] is positioned in this prosperous European EO ecosystem and has three objectives: (i) extracting information and knowledge from big Copernicus data using scalable algorithms, (ii) managing this information and knowledge efficiently, and (iii) integrating it with other data sources to develop demo applications of economic, environmental and societal value.

ExtremeEarth is currently in its final year. Its main achievements so far are the following: (i) two implemented use cases focusing on Food Security and the Polar Regions, (ii) new deep learning architectures for crop type mapping in the context of the Food Security use case, (iii) new deep learning architectures for sea ice mapping in the context of the Polar use case, (iv) the development and open publication of very large datasets for training the deep architectures, (v) scalable semantic technologies for managing, as big linked geospatial data, the information and knowledge extracted from Copernicus data, and (vi) the ExtremeEarth platform that brings all the above technologies together and is used to implement the two use cases.

The rest of the paper presents the above contributions.

2. THE FOOD SECURITY USE CASE

Food Security is a very challenging issue of this century, especially given the changing Earth environment. Irrigation is an important dimension of it requiring reliable water resources either from ground water or from surface water. A large portion of fresh water is linked to snowfall, snow/ice storage and seasonal release. Therefore, water availability maps are an important EO-based product that can support farmers in decision making and irrigation information management.

[1] http://earthanalytics.eu/

doi:10.2760/125905

The goal of the Food Security use case of ExtremeEarth is to *develop high resolution water availability maps* for agricultural areas, allowing a new level of detail for wide-scale irrigation support for farmers [11]. The Danube river basin is the area where the results of the use case have been demonstrated so far. This area was selected for the following reasons: (i) variability in water supply due to changing precipitation patterns leading to extremes events (floods and droughts), (ii) significant portion of irrigated agriculture, (iii) significant water supply from water storage by snow in the Alps, (iv) large interest of demo users, and (v) strong economic, environmental and societal value.

The first stage of this use case was the collection of user requirements during a workshop which was organized by VISTA in Munich in March 2019. The user requirements drove the design and implementation of the Food Security use case which is shown graphically in Figure 1.

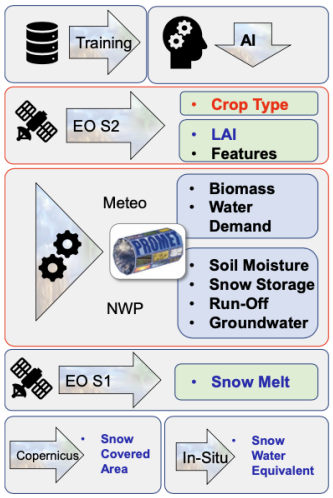

Fig. 1: The Food Security use case

The implementation of the use case draws on the following information: (i) crop type and leaf area index computed using Sentinel-2 images, (ii) biomass, water demand, soil moisture, snow storage, snow run off and groundwater computed using the proprietary land surface modelling software PROMET of VISTA, (iii) snowmelt from Sentinel-1 data, (iv) snow cover products from the Copernicus CryoLand service, and (v) snow water equivalent from in-situ sensors.

The outputs of the use case are field specific irrigation recommendations for specific demo applications in Austria, Hungary and Romania. These consist of recommendations regarding when and how much to irrigate, and yield forecasts with and without optimized irrigation plans.

The implementation of the processing chain of the Food Security use case has been done in the Food Security TEP using the ExtremeEarth platform (see Section 7 and [3]). The deep learning algorithms used for crop type mapping are discussed in Section 4. The semantic technologies that are used are discussed in Section 6.

3. THE POLAR USE CASE

The anticipated economic development of the Arctic, partially driven by reductions in sea ice cover, will see an increase in maritime shipping activity. High quality, timely and reliable information about sea ice and iceberg conditions is vital to ensure that vessels can navigate efficiently and safely with minimal risk to the environment. This information is required by vessels in many sectors, including cargo transport, fisheries, tourism, research vessels, resource exploration and extraction, destination shipping and national coast guard vessels.

The goal of the ExtremeEarth Polar use case is to *produce high resolution ice charts* from massive volumes of heterogeneous Copernicus data. The first stage of the use case was the collection of user requirements during the user workshop of March 2019. Two key technical requirements that resulted from this workshop were: (i) SAR data (Sentinel-1 and other third party missions) were considered the most reliable source of information for the use case, since they are already used widely for operational sea ice charting, and (ii) automatic products to be produced by ExtremeEarth had to maintain the high resolution of this data and the ice charts derived from it (300 meters or better). The technical requirements drove the design and implementation of the Polar use case which is shown graphically in Figure 2.

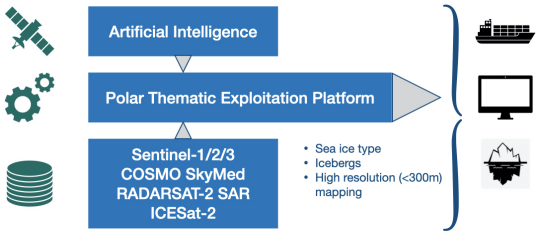

Fig. 2: The Polar use case

The implementation of the use case draws on the following information: (i) Level-1 Sentinel-1 images, (ii) training data compiled manually by expert ice analysts from a variety of sources including other satellite data such as Sentinel-2 and -3 visible and infrared optical, COSMO SkyMed and RADARSAT-2 SAR, and ICESat-2 sea ice freeboard, and in addition shipboard observations from Ice Watch[2].

The outputs of the use case are sea ice concentration and type maps, displaying stages of development (in accordance with the World Meteorological Organization Sea Ice Nomenclature), including fraction of leads and ridges, over the Polar Regions, at a resolution of 300 meters or better.

The implementation of the processing chain of the Polar use case has been done in the Polar TEP [2] using the ExtremeEarth platform (see Section 7 and [3]). The deep learning algorithms used for sea ice classification are discussed in Section 5. The semantic technologies that are used are discussed in Section 6.

[2] https://icewatch.met.no

doi:10.2760/125905

4. DEEP LEARNING FOR CROP TYPE MAPPING

The determination of crops using satellite images is an important component of the pipeline of the Food Security use case discussed in Section 2. For this task, University of Trento developed a deep neural network architecture for crop type mapping using Sentinel-2 image time series [13]. This classification task presents many challenges: (i) the considered time series are noisy, due to the presence of clouds that corrupts the multi-temporal spectral signature, thus affecting the classification results, (ii) time series of different tiles are made up of images acquired in different dates (different temporal sampling), and (iii) a large training dataset of labeled samples is needed to train the deep model.

To address these challenges, the methodology of [13] consists of three main steps: (i) a preprocessing step that generates temporally homogeneous time series of images across tiles that accurately represent the phenological behavior of the crops, (ii) an extraction step that automatically establishes a large training dataset leveraging publicly available crop type maps based on farmer declarations in a large area of Austria, and (iii) a multi-temporal deep learning classification algorithm based on a Long Short Term Memory neural network. The proposed approach achieves more balanced classification results compared to existing state-of-the-art methods obtaining a mean F1 score of 78.32% and an overall accuracy of 85.86%. The approach of [13] has recently been implemented in Hopsworks (see Section 7) and has been deployed in the Food Security TEP.

An important contribution of ExtremeEarth in this context is the development of the training dataset mentioned above which consists of around 1 million pixels of 16 Sentinel-2 images located in Austria, where each pixel is labelled with one of 13 crop types. The dataset will soon be available in the web site of the project.

5. DEEP LEARNING FOR SEA ICE CHARTING

The core of the Polar use case of ExtremeEarth is sea ice classification. For this task, UiT, KTH and DLR have developed multiple deep neural network architectures (LDA, CNNs, variational auto-encoders, GANs, etc.) described in more detail in [5, 6, 7]. Some of these architectures have been implemented in Hopsworks (see Section 7) and have been deployed in the Polar TEP.

An important contribution of ExtremeEarth in this context is the development of three training datasets for sea ice classification: (i) A training dataset consisting of 63,048 patches of 30 Sentinel-1 images located in the European Arctic where each patch is labelled with one of 6 ice types. This dataset was developed by expert photo-interpretation and it was used to train three of the CNNs. (ii) A training dataset consisting of around 62 million patches of 24 Sentinel-1 images located in the Belgica Bank of the Greenland Sea, where each patch is labelled with one of 11 ice types. This dataset was developed using active learning and it was used to train the LDA model and one of the CNNs. (iii) A training dataset consisting of 18,000 patches of 12 Sentinel-1 images located in the Danmarkshavn (East coast of Greenland), where each patch is labelled with one of 2 classes (ice or water). This dataset was developed by expert photo-interpretation and it was used to train one of the CNNs.

The first and the third of the above datasets are publicly available on the web site of the project[3] and the same will be true for the second one very soon.

To advance the international state of the art in this area, ExtremeEarth also organized a workshop on "Machine learning for operational sea ice charting" during ESA's Φ-week 2020.

6. BIG DATA TECHNOLOGIES

The previous sections presented the two use cases of ExtremeEarth and the deep learning algorithms deployed in these use cases. The other technical dimension of the project, which is important in the development of the two use cases, is the utilization of linked data technologies that scale to large volumes of heterogeneous geospatial data available in geographically dispersed data sources. To tackle this important challenge, University of Athens and Demokritos have developed the following big data systems:

- GeoTriples-Spark, a scalable implementation of GeoTriples [8] on top of Apache Spark for transforming geospatial data from their legacy formats (e.g., shapefiles) into RDF.
- JedAI-spatial, a scalable system for interlinking RDF data sources by discovering topological relations among geographic features present in these sources [12].
- Strabo 2, a scalable geospatial RDF store developed using Apache Spark and Apache Sedona.
- A scalable extension of the system SemaGrow [1] for federating geospatial data sources.

To evaluate Strabo 2 and SemaGrow, the same partners have developed and published two benchmarks: Geographica 2 [4] and GeoFedBench [14].

All of the above systems are deployed in the two use cases. In both use cases, information and knowledge extracted from satellite images (e.g., crop type maps) together with data from auxiliary data sources are encoded in RDF using the ontology of the relevant use case. Then, the use case is implemented using the above big data systems. For example, in the Food Security use case, we use an ontology to model data sources such as water availability, crop conditions and irrigation information (see Section 2). The ontology also integrates these data sources with the results of the deep learning algorithms and the PROMET model, so that we can provide irrigation recommendations for specific crop fields in an area of interest.

Another example of the use of the above linked geospatial data technologies in ExtremeEarth is [10], where we show how

[3] http://earthanalytics.eu/datasets.html

doi:10.2760/125905

to use geospatial interlinking algorithms, such as the ones implemented in JedAI-spatial, to produce automatic workflows for combining in-situ observational data with satellite images. For the Polar use case, this has been done using observations from the Ice Watch system of MET Norway, which collects data from ships performing visual sea ice observations while navigating the Arctic. This in-situ observational data record the time, point locations, and other important properties of sea ice. Interlinking these observations with satellite images has enabled MET Norway to validate and improve the interpretation of satellite images, improve routine ice charts, and assist in building deep learning algorithm training datasets.

7. THE EXTREMEEARTH PLATFORM

The ExtremeEarth platform brings together the deep learning architectures and the big data technologies presented above and applies them to the development of the two use cases.

The platform is based on Hopsworks, a data intensive AI platform from Logical Clocks. Hopsworks[4] is an open-source framework for the development and operation of machine learning models, available as a managed platform on AWS and Azure and self-managed (open-source or Enterprise version). It has certain unique features that makes it appropriate for the development of deep learning algorithms for EO data: it provides tools to build end-to-end machine learning pipelines, a feature store, management of machine learning artifacts and assets such as experiments and models, first-class support for popular open-source machine learning frameworks such as TensorFlow, PyTorch, Keras and Scikit-Learn, integration with data science tools such as Jupyter notebooks, and infrastructure monitoring functionalities. Hopsworks provides a horizontally scalable platform for deep learning with GPUs and SDKs for hyper-parameter tuning and elastic model serving.

ExtremeEarth has demonstrated that Hopsworks is an excellent platform for developing the two use cases using the big linked geospatial data systems presented above, as it offers a convenient collaborative environment for building data pipelines. For example, a user can import a specific dataset in a project, transform it into RDF and securely share the results with specific other users or projects, who then can perform further processing, such as interlinking or querying. Hopsworks supports dynamic roles for users accessing and processing such datasets, which enables data owners to securely give access to datasets in a project, knowing the data cannot be exported outside the project or cross-linked with other data sources outside the project. This security model is built on TLS certificates and enables Hops[5] to operate as the only multi-tenant Hadoop platform. In order to perform these tasks, users and developers only need to interact through the human-usable interface of the platform, that offers ready-to-use deployments of popular cloud data storage and processing tools like Apache Hive, Apache Spark and Apache Kafka. Also, using this interface the users can collaborate in order to specify and execute their data pipeline in a Jupyter Notebook and effortlessly monitor the execution progress and inspect the results.

Finally, we have shown that by implementing the big data systems of Section 6 using Hopsworks, we can outperform competitor systems and scale to TBs of geospatial data [9].

[4] https://www.logicalclocks.com/

[5] https://hopsworks.readthedocs.io/en/stable/overview/introduction/what-hops.html

8. SUMMARY

We gave an overview of ExtremeEarth and its main contributions up to today. As the project reaches its conclusion, the ExtremeEarth team is working on the following problems: validation of the deep learning models, detailed experimental evaluation of the implemented big linked geospatial data systems using Geographica 2 and GeoFedBench, and integrating all available technologies to build demos of the two use cases in the Food Security and Polar TEPs.

REFERENCES

[1] A. Charalambidis, A. Troumpoukis, and S. Konstantopoulos. Semagrow: optimizing federated SPARQL queries. In *SEMANTICS*, 2015.

[2] A. Everett, A. Marinoni, A. Cziferszky, D. Arthurs, D.-A. Pantazi, G. Stamoulis, G. Mandilaras, J. L. Pedersen, M. Datcu, N. Hughes, P. Wagner, S. Khaleghian, T. Kræmer, and Å. Kiærbech. Implementation and evaluation of the Polar use case-v1. ExtremeEarth Deliverable D5.3, Available from http://earthanalytics.eu/deliverables.html, 2020.

[3] D. H. Hagos, T. Kakantousis, V. Vlassov, S. Sheikholeslami, T. Wang, J. Dowling, A. Fleming, A. Cziferszky, M. Muerth, F. Appel, D.-A. Pantazi, D. Bilidas, G. Papadakis, G. Mandilaras, G. Stamoulis, M. Koubarakis, A. Troumpoukis, and S. Konstantopoulos. Software architecture for Copernicus Earth observation data. In *BiDS*, 2021.

[4] T. Ioannidis, G. Garbis, K. Kyzirakos, K. Bereta, and M. Koubarakis. Evaluating geospatial RDF stores using the benchmark Geographica 2. *Journal on Data Semantics*, 2021.

[5] C. Karmakar, C. O. Dumitru, G. Schwarz, and M. Datcu. Feature-free explainable data mining in SAR images using latent Dirichlet allocation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:676–689, 2021.

[6] S. Khaleghian, T. Kræmer, A. Everett, Åshild Kiærbech, N. Hughes, T. Eltoft, and A. Marinoni. Synthetic aperture radar data analysis by deep learning for automatic sea ice classification. In *The European Conference on Synthetic Aperture Radar*, 2021.

[7] T. Kræmer, S. Khalegian, A. Marinoni, C. Dumitru, M. Datcu, and T. Eltoft. Deep architectures implementation for the Polar use case-v1. ExtremeEarth Deliverable D2.3.

[8] K. Kyzirakos, D. Savva, I. Vlachopoulos, A. Vasileiou, N. Karalis, M. Koubarakis, and S. Manegold. GeoTriples: Transforming geospatial data into RDF graphs using R2RML and RML mappings. *J. Web Sem.*, 2018.

[9] G. Mandilaras and M. Koubarakis. Transforming big geospatial data into linked data. In *Forthcoming*, 2021.

[10] G. Mandilaras, D.-A. Pantazi, M. Koubarakis, N. Hughes, A. Everett, and Å. Kiærbech. Ice monitoring with ExtremeEarth. In *2nd Workshop on Large Scale RDF Analytics*, 2020.

[11] S. Migdall, S. Dotzler, C. Miesgang, F. Appel, M. Muerth, H. Bach, G. Weikmann, C. Paris, D. Marinelli, and L. Bruzzone. Water stress assessment in Austria based on deep learning and crop growth modelling. In *Submitted to BiDS*, 2021.

[12] G. Papadakis, G. Mandilaras, N. Mamoulis, and M. Koubarakis. Progressive, holistic geospatial interlinking. In *The Web Conference*, 2021.

[13] C. Paris, G. Weikmann, and L. Bruzzone. Monitoring of agricultural areas by using Sentinel 2 image time series and deep learning techniques. In *SPIE Remote Sensing Conference*, 2020.

[14] A. Troumpoukis, S. Konstantopoulos, G. Mouchakis, N. Prokopaki-Kostopoulou, C. Paris, L. Bruzzone, D.-A. Pantazi, and M. Koubarakis. GeoFedBench: A benchmark for federated GeoSPARQL query processors. In *ISWC*, 2020.

doi:10.2760/125905