





SURVEY PAPER

Earth System Data Cubes: Avenues for advancing Earth system research

David Montero^{1,2,3} , Guido Kraemer^{1,2} , Anca Angheloa⁴, César Aybar^{5,6}, Gunnar Brandt⁷, Gustau Camps-Valls⁵, Felix Cremer⁸, Ida Flik^{1,2}, Fabian Gans⁸, Sarah Habershon^{1,2}, Chaonan Ji^{1,2}, Teja Kattenborn⁹, Laura Martínez-Ferrer⁵, Francesco Martinuzzi^{1,2,10}, Martin Reinhardt^{1,2} , Maximilian Söchting^{1,2,11}, Khalil Teber^{1,2} and Miguel D. Mahecha^{1,2,3,10,12} 

¹Remote Sensing Centre for Earth System Research (RSC4Earth), Leipzig University, 04103, Leipzig, Germany

²Institute for Earth System Science & Remote Sensing, Leipzig University, 04103, Leipzig, Germany

³German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, 04103, Leipzig, Germany

⁴European Space Research Institute (ESRIN), European Space Agency (ESA), 00044, Frascati, Italy

⁵Image Processing Laboratory, Universitat de València, 46980, València, Spain

⁶Water Competence Center (CCA), 15086, Lima, Perú

⁷Brockmann Consult GmbH, 21029, Hamburg, Germany

⁸Max Planck Institute for Biogeochemistry, 07745, Jena, Germany

⁹Sensor-based Geoinformatics, University of Freiburg, 79106, Freiburg, Germany

¹⁰Center for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI), Leipzig University, 04105, Leipzig, Germany

¹¹Image and Signal Processing Group, Leipzig University, 04109, Leipzig, Germany

¹²Department of Remote Sensing, Helmholtz Centre for Environmental Research (UFZ), 04318, Leipzig, Germany

Corresponding author: David Montero; Email: david.montero@uni-leipzig.de

Received: 02 July 2023; **Revised:** 31 July 2024; **Accepted:** 13 August 2024

Keywords: Earth System Science; Data Cubes; Artificial Intelligence; Data Life Cycle

Abstract

Recent advancements in Earth system science have been marked by the exponential increase in the availability of diverse, multivariate datasets characterised by moderate to high spatio-temporal resolutions. Earth System Data Cubes (ESDCs) have emerged as one suitable solution for transforming this flood of data into a simple yet robust data structure. ESDCs achieve this by organising data into an analysis-ready format aligned with a spatio-temporal grid, facilitating user-friendly analysis and diminishing the need for extensive technical data processing knowledge. Despite these significant benefits, the completion of the entire ESDC life cycle remains a challenging task. Obstacles are not only of a technical nature but also relate to domain-specific problems in Earth system research. There exist barriers to realising the full potential of data collections in light of novel cloud-based technologies, particularly in curating data tailored for specific application domains. These include transforming data to conform to a spatio-temporal grid with minimum distortions and managing complexities such as spatio-temporal autocorrelation issues. Addressing these challenges is pivotal for the effective application of Artificial Intelligence (AI) approaches. Furthermore, adhering to open science principles for data dissemination, reproducibility, visualisation, and reuse is crucial for fostering sustainable research. Overcoming these challenges offers a substantial opportunity to advance data-driven Earth system research, unlocking the full potential of an integrated, multidimensional view of Earth system processes. This is particularly true when such research is coupled with innovative research paradigms and technological progress.

Impact Statement

Today, we have the capability to continuously monitor a broad array of processes within the Earth system at high spatio-temporal resolutions. However, it is only through combined analysis that these data reveal their full potential in advancing Earth system research. Earth System Data Cubes (ESDCs) possess transformative potential in this regard, yet they are accompanied by significant challenges throughout their life cycle. This paper offers a detailed exploration of these challenges, highlighting the importance of rigorous ESDC analysis while warning against potentially misleading outcomes from naive applications.

1. Introduction

Humanity possesses the capability to observe and model the majority of Earth's subsystems, generating vast amounts of data with unprecedented resolution, quality, and coverage (Simmons et al., 2016; Peng et al., 2021; Bauer et al., 2021a). The co-interpretation of these diverse datasets represents an unprecedented opportunity for understanding the intricacies of the Earth system (Runge et al., 2019; Mahecha et al., 2020; Tuia et al., 2023). However, this wealth of heterogeneous data comes with substantial challenges. The sheer volume of data, characterised by variations in spatial and temporal resolution as well as data curation levels, coupled with the high complexity of processes encoded in these multi-dimensional datasets, renders conventional data processing and interpretation methods unsuitable (Boulton, 2018; Sudmanns et al., 2020).

Recognising the need for a simple yet robust data infrastructure to facilitate Earth system data interoperability led to the emergence of various data cube concepts (Nativi et al., 2017; Baumann et al., 2019; Giuliani et al., 2019; Kopp et al., 2019; Mahecha et al., 2020, and others). We refer to Earth System Data Cubes (ESDCs) as frameworks where diverse datasets are integrated into a unified, highly interoperable system, organised on a common spatio-temporal grid (a more formal definition is given in Section 2.1). The essence of ESDCs is to convert the vast array of Earth system data into readily accessible data streams, apt for a variety of Earth system research domains. Such frameworks are gaining widespread acceptance in Earth system research as a solution for managing complex Earth Observation (EO) data.

Given the simplicity of such structures, various initiatives have greatly enhanced the use of EO data derived from satellite remote sensing and other large-scale array data, such as climate model outputs. Initiatives building on an ESDC concept originally developed their data in hand-crafted ways (e.g. Mahecha et al., 2020; Estupiñan Suarez et al., 2021; Walther et al., 2022) or created systems supporting on-demand generations of ESDCs (e.g. Appel and Pebesma, 2019; Killough, 2018; Schramm et al., 2021). Earth system data providers have invested tremendous efforts in compiling extensive data catalogues, which can be used for the development of further ESDCs. Notable examples of such catalogues are provided by Google Earth Engine (GEE, Gorelick et al., 2017)¹, Microsoft Planetary Computer², or the Open Geospatial Data Catalogue of Amazon Web Services (AWS)³. Additionally, there is a constant effort to increase the adoption of ESDCs (including generation and analysis) within cloud environments (Zellner et al., 2024). Therefore, ESDCs can be efficiently generated and used in virtual laboratories, such as the DeepESDL (Brandt et al., 2023; Sturm, 2023)⁴, or the agricultural virtual lab⁵.

This access to straightforward aligned Earth system data has facilitated numerous Earth system research questions. For instance, researchers have employed both linear and non-linear dimensionality reduction methods to generate global indicators for the terrestrial biosphere (Kraemer et al., 2020), uncover the main modes of Earth system variables (Bueso et al., 2020), quantified spatial dynamics of vegetation responses to ENSO in South America (Estupinan-Suarez et al., 2023), or gained major insights

¹ <https://developers.google.com/earth-engine/datasets/>

² <https://planetarycomputer.microsoft.com/catalog>

³ <https://aws.amazon.com/earth/>

⁴ <https://deepesdl.readthedocs.io/>

⁵ <https://agriculturevlab.eu/>

on Land Use and Cover Change (LUCC, Santos et al., 2019). Specifically, EO data cubes, or ESDCs comprising satellite remote sensing imagery, have been instrumental in applications such as learning the vegetation response to climate drivers using Recurrent Neural Network (RNN) architectures (Martinuzzi et al., 2023), quantifying drought legacy effects on gross primary production (Yu et al., 2022), and detecting spatio-temporal extreme events (Mahecha et al., 2017).

However, if the goal is for ESDCs to evolve and become sustainable data infrastructures, it is essential to develop robust ESDC life cycles. Considering the unique characteristics of ESDCs, we cannot merely apply existing research data life-cycle concepts; instead, we must identify and address the peculiarities specific to ESDCs. It is necessary to create opportunities for continuous improvement and to address current challenges by leveraging contemporary technological advancements, specifications, and research paradigms. For instance, data formats and sharing protocols must evolve to align with the current status of cloud-based technologies and standards, in accordance with the adoption of FAIR Open Science principles (Wilkinson et al., 2016). Moreover, transforming heterogeneous data into an analysis-ready format aligned with a multidimensional spatio-temporal grid is often complex and subject to application-specific variations (Giuliani et al., 2019; Zuefle et al., 2021).

The resulting data format, though straightforward and relatively easy to analyse, encompasses inherent complexities (Béjar et al., 2023). These intricacies necessitate careful consideration during data analysis, requiring a profound understanding of the nature of Earth system processes. Naive analyses based on ESDCs can potentially lead to misleading interpretations as pointed out, e.g. by Meyer et al. (2018); Rußwurm et al. (2023) or Sweet et al. (2023). Common pitfalls include model performance inflation caused by spatio-temporal auto-correlation, biased sampling, and inaccurate spatial aggregations. It's only by adequately addressing these challenges that the full potential of ESDCs can be realised, aligning with the perspectives of various authors (Reichstein et al., 2019; Irrgang et al., 2021; Hsieh, 2022; Sun et al., 2022; Persello et al., 2022; Tuia et al., 2023). Topics widely discussed today are generative processes in Artificial Intelligence (AI) that could enable researchers to reconstruct unseen data (Rüttgers et al., 2019; Oyama et al., 2023). Another promising direction is the potential for making causal inferences solely from data (Runge et al., 2019; Krich et al., 2021; Christiansen et al., 2022; Camps-Valls et al., 2023). Also, integrating physical constraints and domain knowledge in the inference process can lead to more plausible semi-empirical predictions (Ilie et al., 2017; Karniadakis et al., 2021; Camps-Valls et al., 2021; Cortés-Andrés et al., 2022). Concurrently, advances in data processing and visualisation technologies not only enhance data exploration and analysis but also aid in disseminating research findings (Söchting et al., 2023).

This paper seeks to identify the challenges inherent in the complete ESDC life cycle while, at the same time, highlighting the potential to advance Earth system research through these data structures. The manuscript is organised as follows: **Section 2** introduces the concept of ESDC and its relationship to information-preserving systems for Earth system data. In **Section 3**, we elaborate on the ESDC life cycle, displaying the obstacles encountered during data processing and proposing pathways toward creating analysis-ready ESDCs. **Section 4** explores the transformative possibilities stemming from contemporary AI advancements in Earth system research while **Section 5** cautions against the risks of uninformed ESDC analysis. **Section 6** addresses the technical facets of manipulating ESDCs throughout their life cycle, offering insights into technologies that can streamline Earth system data processing. Lastly, in **Section 7**, we examine the challenges associated with data visualisation in the context of ESDCs. Through this paper, we aim to outline the complexities and opportunities associated with employing ESDCs, hopefully paving the way for advancements in Earth system research.

2. The Art of Data Cubes

Data cubes are renowned for their capacity to serve as multidimensional arrays of data, enabling the representation of values across various dimensions of interest within a specific domain. Specialised data cubes designed for analytical queries in database systems, such as Online Analytical Processing (OLAP, Chaudhuri and Dayal, 1997) cubes, have been integrated with Geographical Information System (GIS)

databases to give rise to Spatial OLAP (SOLAP, Rivest et al., 2005) cubes. SOLAP infrastructures, although traditionally associated with vector data, are also available for raster data (Kasprzyk and Donnay, 2017). Database systems have proven effective in storing and managing Earth system data in the form of data cubes, exemplified by array database solutions like Rasdaman (Baumann et al., 1998). Additionally, data cube infrastructures can be employed to store indexed files (Killough, 2018), thus safeguarding the information that might otherwise be lost during data transformation processes, such as reprojection. Here, we rely on a specific interpretation of data cubes, specifically tailored to tackle the vast volumes and interoperability challenges of Earth system data. We first explain the concept of ESDCs, but also provide an overview of related information-preserving structures, namely image collections and information-preserving data cubes, showcasing how they interface with ESDCs.

2.1. What are Earth System Data Cubes (ESDCs)?

The concept of ESDCs was introduced along with the Earth System Data Lab (ESDL, Mahecha et al., 2020), an integrated data and analytical hub that aimed to unify multiple heterogeneous Earth system data streams into a standard data model with a unique Coordinate Reference System (CRS). ESDCs represent multidimensional data structures designed to facilitate streamlined access, analysis, and manipulation of Earth system data. ESDCs comprise labels as **dimensions** defining the cube's axes, an array of **grids** with their associated coordinate values distributed along these dimensions, and univariate **data** associated with each grid cell. Furthermore, in this paper, we add a new component: a suite of **attributes** that characterise the data, the dimensions, and the complete ESDC entity.

The **dimensions** are a set of labels describing the axes of the ESDC. Generally, these dimensions comprise space (e.g. “x” and “y”), time, and variables. Nevertheless, further dimensions can be added (e.g. “pressure levels”, “model ensembles” or “time series components”). It is crucial to emphasise that while ESDCs conventionally incorporate spatial and temporal dimensions (e.g. latitude, longitude, and time), they are not confined to this paradigm (cf. Table 1 of Mahecha et al., 2020). ESDCs can exhibit different dimensions, and the number of dimensions is called the order of the ESDC. Thus, an increment in the ESDC's complexity according to its dimensions is given by their order (e.g. a spatio-temporal grid of a univariate ESDC has an order of 3, while the order of a multivariate ESDC is 4).

The grouping of **grids** consists of discrete subsets derived from the domain of each dimension's axis. The values of these subsets are referred to as coordinates, and, in the case of a regular grid, they determine the data's resolution along that specific dimension. For instance, a grid determining a resolution of 0.5 degrees for the “latitude” dimension in a global ESDC may have a set of coordinates $\text{grid}(\text{latitude}) = \{-89.75, -89.25, \dots, 89.25, 89.75\}$. While coordinates are often associated with numerical values (e.g. latitudes and longitudes), they can encompass a wide range of values. For instance, timestamps in a “time” dimension with a set of coordinates $\text{grid}(\text{time}) = \{“2022 - 01 - 01”, “2022 - 01 - 02”, \dots, “2022 - 12 - 31”\}$, or components derived from a time series decomposition approach in a “component” dimension with a set of coordinates $\text{grid}(\text{component}) = \{“raw”, “trend”, “seasonal”, “residual”\}$. The grids within an ESDC exhibit the following characteristics: 1) In the case of spatial dimensions, they reference the same CRS, 2) the coordinates within a grid share identical units, and 3) they must consist of at least two coordinates; otherwise, the dimension (and consequently the grid) is omitted. It's important to note that, given these properties, irregular grids are also possible, with the temporal dimension grid being a typical example in EO data due to the irregular revisit times of some satellite missions (e.g. Sentinel-2).

The array of **data** represents scalar values corresponding to each grid cell. Typically, the data spans from observed measurements to modelled values. Nevertheless, one can also encounter higher-order features (that is data derived from operations performed on the original values), such as outcomes from time series decomposition or AI-generated products. Furthermore, flag values, which delineate data status, can be incorporated. Cells without data are denoted as “NA” (that is not available).

The collection of **attributes** comprises a series of key-value objects that provide additional details about the data. These objects serve as metadata and can offer descriptions ranging from individual

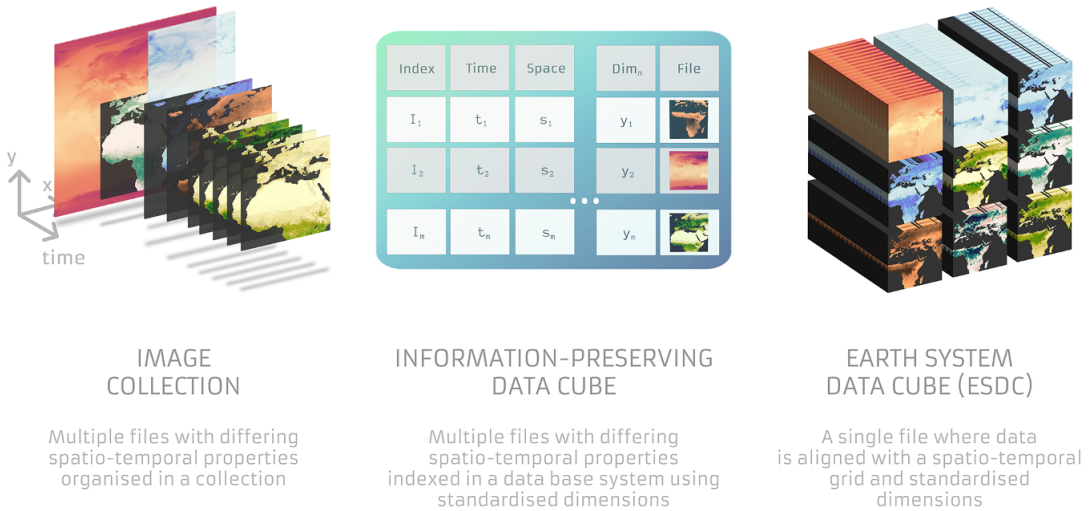


Figure 1. Representations of different storage systems for gridded data in Earth system research: Image collections (left), information-preserving data cubes (centre), and Earth system data cubes (ESDCs, right). Differences in these abstract representations have deep implications for data storage systems, accessibility, interoperability and metadata definitions.

variables (including their associated dimensions) to the entire ESDC. The information contained within these attributes typically encompasses a wide range of elements, including, but not limited to, names, acronyms, units, flag definitions, versions, and source details.

2.2. Relation of ESDCs to Image Collections and Data Cubes

Earth system data often exhibits heterogeneity and irregularity, particularly within EO data. Variability can manifest in different spatial resolutions, time units, projections, formats, and more, sometimes even within the same data product. Consequently, two robust approaches to retaining data integrity without succumbing to information loss due to transformations (e.g. reprojection, reduction, and resampling) are to utilise **image collections** (refer to Appel and Pebesma, 2019 for a comprehensive distinction between image collections and conventional data cubes) or to adopt a process where original files are stored and indexed within a **information-preserving data cube** infrastructure based on their file metadata (Figure 1). In the latter approach, the original files can be stored locally or in the cloud while preserving the essential information intact.

A successful example of **image collections** is the GEE Catalogue. This extensive, multi-petabyte catalogue stores data in tiled images, where each image may encompass multiple bands, thereby preserving essential information. Furthermore, these images can be organised into an image collection if they share relevancy. GEE also offers the computational resources necessary for accessing and analysing their catalogued data. Within GEE, data cube-like operations can be seamlessly executed through dynamic on-the-fly reprojection, resampling, and reduction for the tiles where a subset of pixels was explicitly requested. It is worth noting, however, that users are required to conform to the specific Application Programming Interfaces (API) provided by GEE for processing and analysing the data effectively.

Standardising image collections and their access brings simplicity and promotes data usage across platforms. Currently, a widely recognised standard is the Spatio-Temporal Assets Catalog (STAC) specification. This specification empowers users to query data assets based on metadata and spatio-temporal criteria. Coupled with domain-specific API clients available for multiple programming

languages (cf. Section 6.2) and GIS software (e.g. QGIS STAC Plugin⁶), users can easily retrieve data. The flexibility of the STAC specification has prompted numerous data providers to adopt it for creating their own data catalogues⁷, with notable examples including Microsoft Planetary Computer Catalogue⁸ and the United States Geological Survey (USGS) Landsat Archive Catalogue (stored in the Amazon Simple Storage Service, S3)⁹.

The **information-preserving data cube** approach is exemplified by the Open Data Cube (ODC) initiative, a prominent model in this field (Killough, 2018; Killough et al., 2020)¹⁰. This approach has played a pivotal role in informing governmental actions and policies, as evidenced by their integration into national and regional data cube frameworks (Dhu et al., 2019; Sudmanns et al., 2022). Noteworthy instances of these initiatives include Digital Earth Africa (DE Africa, formerly known as Africa Regional Data Cube, Killough, 2019), Digital Earth Australia (DE Australia, previously Australian Geoscience Data Cube, Lewis et al., 2017; Dhu et al., 2017), the Colombian Data Cube (CDCol, Ariza-Porras et al., 2017; Bravo et al., 2017; Villamizar et al., 2018), and the Swiss Data Cube (SDC, Giuliani et al., 2017).

While both of these approaches excel in preserving data integrity and offering flexibility for various analyses, achieving Earth system data interoperability necessitates their integration into a unified structure through ESDCs. These ESDCs can be constructed from either approach. For instance, in the case of image collections, it is feasible to request pixels from GEE (Clinton, 2023), and data transformations can be executed within the GEE cloud-based environment before downloading the data. It's important to note that limitations related to the size of the requested data can be a potential concern in this process. Alternatively, STAC simplifies the process, particularly when combined with cloud-ready formats. This lazily enables the creation of ESDCs. In the information-preserving data cube approach, platforms like ODC offer a comprehensive system for transforming original data into ESDCs and even provide mechanisms for storing the resulting ESDCs within the data cube infrastructure¹¹. Noteworthy is openEO (Schramm et al., 2021), an API striving to connect multiple backends in a standardised way, including image collection providers (e.g. GEE) and information-preserving data cubes (e.g. ODC)¹², to generate ESDCs.

3. The ESDC Life cycle

Creating an ESDC from multiple sources, including source files, data cubes, or image collections, is a multifaceted process. The ESDC life cycle, as illustrated in Figure 2, encompasses several crucial stages, each playing a vital role in the generation, analysis, and effective utilisation of these data structures. The ESDC life cycle comprises the following key phases: data collection, curation, cubing, harmonisation, transformation, analysis, and reuse. These phases are linked, reflecting the meticulous efforts involved in ESDCs' development. In parallel to these stages, metadata generation occurs concurrently with data transformations, data exploration, visualisation, and dissemination. This section provides an overview of the ESDC life cycle, emphasising relevant considerations that contribute to the streamlined development and utilisation of ESDCs.

3.1. Collection

Given that data providers frequently utilise diverse formats and protocols for data sharing, particularly in proportion to the multidimensional complexity of the data, the establishment of streamlined access mechanisms becomes imperative. Traditionally, File Transfer Protocol (FTP) servers have been used for

⁶ <https://github.com/stac-utils/qgis-stac-plugin>

⁷ <https://stacindex.org/catalogs/>

⁸ <https://planetarycomputer.microsoft.com/>

⁹ <https://www.usgs.gov/core-science-systems/nli/landsat/landsat-commercial-cloud-data-access>

¹⁰ <https://www.opendatacube.org/ceos>

¹¹ <https://www.opendatacube.org/overview>

¹² <https://openeo.org/software.html>

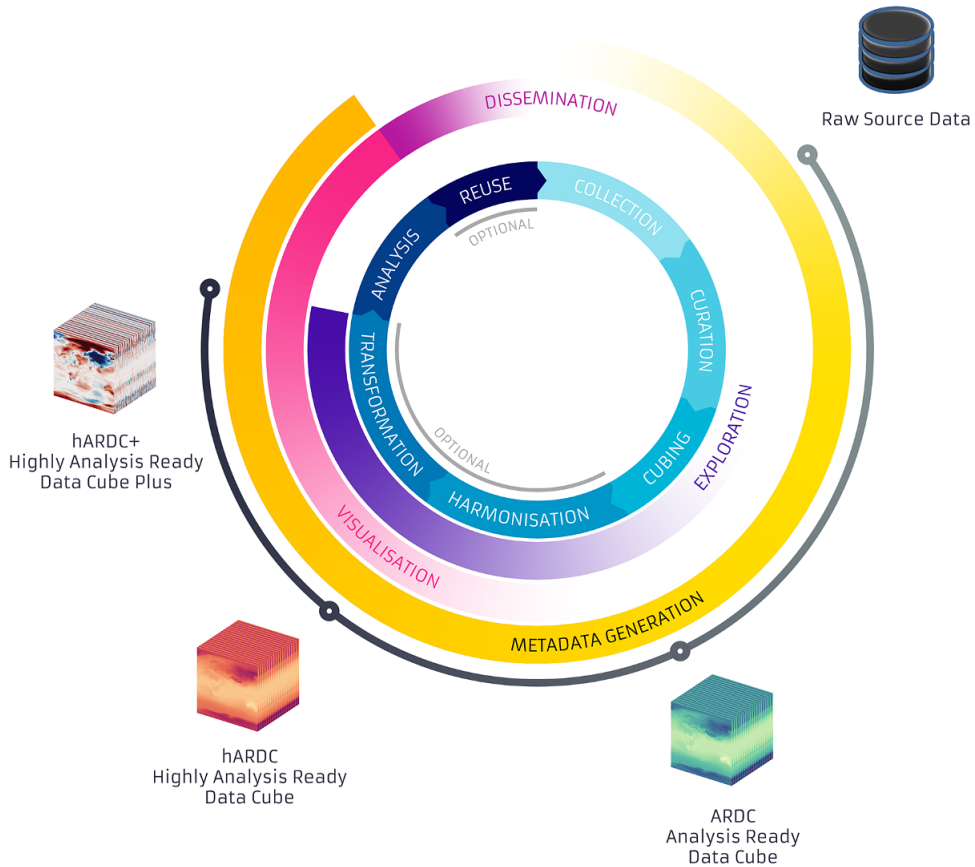


Figure 2. ESDC life cycle. The inner circle represents data processing tasks, and the outer circles represent ancillary tasks that run parallel to the processing steps, involving activities such as data exploration, visualisation, dissemination, and metadata generation. The outermost circle of the diagram illustrates the readiness level of the processed ESDCs at specific points within the cycle.

data sharing. However, to enhance data discoverability and usability, data providers are increasingly adopting data stores that offer persistent and standardised data storage. Repositories play a vital role in this process by standardising metadata, enabling easy search and retrieval of assets through metadata queries. Recently, more and more data providers offer APIs to facilitate efficient querying of metadata and access to the data itself by adopting specifications such as STAC and enabling range requests for cloud-optimised data (e.g. Zenodo recently started to support HTTP range requests¹³).

The flexibility of these specifications enhances data interoperability by enabling the development of extensions that simplify data integration. For instance, the Electro-Optical STAC-extension¹⁴ has been created to facilitate the integration of multispectral remote sensing data by expanding the capabilities of STAC to accommodate specific requirements and metadata associated with this kind of data. Looking ahead, the advantages of data interoperability may potentially extend beyond the realm of raw source data, encompassing entire ESDCs. The datacube STAC-extension¹⁵ (currently in a “candidate” Extension

¹³ <https://blog.zenodo.org/2021/12/07/2021-12-07-hardening-our-service/>

¹⁴ <https://github.com/stac-extensions/eo>

¹⁵ <https://github.com/stac-extensions/datacube>

Maturity level) has been developed with the primary objective of advancing the integration and interoperability of structured data representations like ESDCs within the STAC ecosystem. This effort aims to broaden the scope of opportunities for reusing ESDCs in new data processing pipelines.

Additionally, the efficiency of data access and collection is contingent upon data formats. GeoTIFF is arguably the most used and renowned data format for georeferenced raster data. This format adds a standard specification for the TIFF format that describes the spatial properties of the raster. It is widely used for EO products such as Landsat imagery. The need to operate in cloud environments has driven the development of cloud-optimised geospatial data formats. Consequently, the GeoTIFF format has evolved to the Cloud-Optimised GeoTIFF (COG)¹⁶ format, enhanced to function efficiently in cloud environments through HTTP range requests. COGs offer several advantages over traditional GeoTIFFs, including reduced latency in data retrieval, faster visualization of large datasets, and a tiled structure that enables parallel processing. The significance of this format is underscored by its recent approval as an Open Geospatial Consortium (OGC) standard^{17,18}.

When the dimensionality of the data increases, formats such as NetCDF or HDF5 are typically used to encapsulate data and coordinate values. Tiling and chunking allow efficient access to big data arrays for both data formats. However, these formats are not inherently optimised for cloud environments. The Zarr specification¹⁹ addresses this limitation and can be used directly in cloud environments, offering several advantages over NetCDF and HDF5. Zarr enables more efficient chunk access for parallel processing, provides better support for distributed computing, and offers improved read and write speeds, particularly in cloud storage systems. Moreover, Zarr's flexible chunking scheme allows for optimised data access patterns, and its simpler metadata structure facilitates easier data discovery and management. Additionally, specifications such as geo-zarr²⁰ and the xcube dataset convention²¹ have been introduced to further enhance data interoperability and compatibility within the context of Earth system data.

3.2. Curation

Effective data curation stands as a critical anchor in the preparation of data for subsequent spatio-temporal processes and analysis via ESDCs (Marujo et al., 2022). The transformation of raw data into Analysis-Ready Data (ARD) has emerged as an essential prerequisite across multiple initiatives. ARD ensures that data are readily amenable to queries, analysis, and application development. Notable instances of these initiatives include DE Africa²², DE Australia²³, and the Brazil Data Cube (Ferreira et al., 2020; Marujo et al., 2022), among others.

The Committee on Earth Observation Satellites (CEOS) has precisely defined ARD as “satellite data that have been processed to a minimum set of requirements and organized into a form that allows immediate analysis with a minimum of additional user effort and interoperability both through time and with other datasets”²⁴. In this definition, ARD also exhibit interoperability both across time and with other datasets (refer to Siqueira et al., 2019 for an overview of the CEOS ARD for Land initiative, CARD4L). CEOS has established a comprehensive set of Product Family Specifications (PFS) tailored to various data groups, including Surface Reflectance, Surface Temperature, Polarimetric Radar, and more. These specifications undergo rigorous peer review processes across multiple satellite platforms, such as Landsat and Sentinel collections, to obtain CEOS ARD approval. It's worth noting that there are ongoing efforts to develop additional PFS, including Interferometric Radar and LiDAR Terrain and Canopy Height. It is also

¹⁶ <https://www.cogeo.org/>

¹⁷ <https://docs.ogc.org/is/21-026/21-026.html>

¹⁸ <https://www.ogc.org/press-release/cloud-optimized-geotiff-cog-published-as-official-ogc-standard/>

¹⁹ <https://zarr.dev/>

²⁰ <https://github.com/zarr-developers/geo-zarr-spec>

²¹ <https://xcube.readthedocs.io/en/latest/cubespec.html>

²² <https://www.digitalearthfrfrica.org/platform-resources/analysis-ready-data>

²³ <https://www.dea.ga.gov.au/about/analysis-ready-data>

²⁴ <https://ceos.org/ard>

worth noting that the OGC has recently addressed the CEOS ARD concept by forming a new Standards Working Group (SWG) to define a generic multi-part standard specifying a set of minimum requirements for geospatial products to be considered ARD²⁵.

It is important to recognise that achieving an ARD level can extend beyond minimum standard specifications. Obtaining ARD often involves crucial preprocessing and data curation tasks that are tailored to the unique requirements of the application domain. For instance, in the context of EO data, these tasks may encompass but are not limited to, cloud and cloud shadow masking (refer to Skakun et al., 2022 for a comprehensive intercomparison exercise of multiple cloud and cloud shadow masking methods), snow masking (e.g. Richiardi et al., 2021), and the correction of Bidirectional Reflectance Distribution Function (BRDF) effects to derive Nadir BRDF Adjusted Reflectance (NBAR) values (e.g. Roy et al., 2016).

3.3. Cubing

The concept of ARD may exhibit some subjectivity depending on the specific application. This subjectivity pertains to the data that populates an ESDC. In contrast, ESDCs inherently represent straightforward yet robust analysis-ready integrated entities, capable of simplifying a broad spectrum of analytical tasks (Baumann et al., 2019). An ESDC filled with ARD is often called an **Analysis-Ready Data Cube (ARDC)**, a concept widely employed in DeepESDL. To generate an ARDC, the critical step involves aligning data onto a unified grid. Domain experts predefine this grid, and all data sources must conform. Furthermore, the efficacy of the ARDC processing is significantly influenced by the implementation of an optimal chunking strategy for this grid. This strategy must be tailored to facilitate efficient data processing across diverse analytical scenarios. For instance, analyses focused on temporal dynamics benefit from chunking strategies that preserve the temporal dimension, whereas spatial analyses or cartographic visualisations are optimised by maintaining spatial dimensions within chunks. In scenarios requiring multi-temporal spatial analysis, a hybrid approach combining both temporal and spatial preservation in chunking can be advantageous.

When the grid moves in the spatio-temporal domain, the varying spatio-temporal resolutions and coverage among multiple data sources require selecting adequate methods to fit the data into the predefined grid. Datasets with varying spatial resolutions and coverage must be resampled onto a standard spatial grid. This process often requires modifying the data (Cracknell, 1998). While non-destructive algorithms such as nearest neighbours can preserve data values (at the cost of duplicating or ignoring values), significant differences in spatial resolution often require transformation through (non-) linear resampling methods, such as cubic convolution or advanced fusion techniques (Nicolakopoulos, 2008). Complex AI methods can be employed to perform spatial transformations while preserving the quality of the measured variable (e.g. multi-image super-resolution algorithms, Michel et al., 2022; Razzak et al., 2023). Another often overlooked issue arises when dealing with extensive variables. In such instances, it is crucial to ensure that, for example, mass balances are not distorted in the process of creating new products.

It is important to note that the application of resampling methods, particularly in the context of generating Global ESDCs covering the entire planet (e.g. Mahecha et al., 2020), may introduce geometric distortions. Projecting global datasets onto a plane can distort the data in terms of area, distances, and angles (Snyder and Voxland, 1989), posing challenges for subsequent analysis (cf. Section 5.1). This can be alleviated by using a Discrete Global Grid System (DGGS, Kmoch et al., 2022). This kind of grid system seeks to minimise distortions, harmonise cell sizes and maintain consistent distances from neighbours. Defining standards and solutions for efficient chunk storage, subsetting, and integration into the ESDC framework will be a challenging future task. Still, it could lead to significant improvements in both the performance and accuracy of spatial algorithms.

²⁵ <https://www.ogc.org/press-release/ogc-forms-new-analysis-ready-data-standards-working-group>

In the case of Regional ESDCs (e.g. Estupiñan Suarez et al., 2021), which may cover entire continents, oceans, or administrative regions at various hierarchical levels, selecting an appropriate CRS is crucial to ensure minimal geometric distortion. On local scales, Local ESDCs (also referred to as mini cubes, Requena-Mesa et al., 2021) cover smaller areas of interest (e.g. Walther et al., 2022), ideally characterised by high spatial resolutions ranging from sub-meters to meters. Using local ESDCs together with a local CRS enables to minimise distortions.

When dealing with datasets characterised by varying temporal grids, even if they share the same date-time units, irregular temporal grids may emerge. These discrepancies can introduce temporal gaps within the time dimension. In cases where datasets exhibit varying date-time units, especially when working with datasets featuring finer date-time units (e.g. daily records), it becomes necessary to aggregate them to align with a predefined coarser temporal grid (e.g. monthly records). While this process is straightforward for regularly sampled data, it can pose challenges for EO data with long revisit periods (e.g. Landsat data). These challenges can potentially introduce uncertainties during aggregation. Substantial gaps in EO data can have a detrimental impact on the accuracy and representativeness of the aggregated results. This concern is further exacerbated when additional gaps arise due to data disturbances, such as cloud and shadow interference.

3.4. Harmonisation

Additional post-processing of data variables may be necessary to address Earth system challenges. This entails further data curation to obtain a fully gap-filled, harmonised product with evenly spaced time steps. In alignment with the naming conventions established for EO data cubes by Frantz, 2019, we refer to a thoroughly harmonised ESDC as a **highly ARDC (hARDC)**.

Data harmonisation is crucial to ensure the consistency and compatibility of variables obtained or generated using different methodological or technical approaches (Wulder et al., 2015). When discrepancies exist between data measurement or production methods, it can introduce inconsistencies that hinder subsequent analyses involving the specific variables (Vogeler et al., 2018). To address this, one approach is to create separate variables that represent the same measured quantity, highlighting the differences between them. However, to enhance spatio-temporal resolution and coverage, harmonisation of variables is often necessary (e.g. harmonising reflectance values from Sentinel-2 and Landsat, Claverie et al., 2018; Marujo et al., 2023).

This can be achieved through simple methods that involve sampling data from the same spatio-temporal index in both variables to establish a direct conversion model (e.g. using matched observations to match Landsat 8 and Sentinel-2, Shang and Zhu, 2019). Alternatively, more advanced AI models can harmonise data by incorporating one or more additional variables (e.g. creating a global product of OCO-2 Sun-Induced Fluorescence, SIF, Li and Xiao, 2019). This may require the development of an entire AI pipeline to extend a variable with newly available data or reconstruct it, especially in cases where the variable was not previously measured (e.g. reconstructing SIF from TROPOMI, Chen et al., 2022). In this sense, data harmonisation also encompasses projecting data in simulated future scenarios (e.g. projecting vegetation dynamics for the rest of the century, Mahowald et al., 2016). In addition, it is crucial to incorporate uncertainty metrics to facilitate accurate and reliable future analysis using the harmonised data variables (cf. Section 4.3).

Additionally, to effectively use algorithms that incorporate temporal structures, such as Recurrent Neural Networks (RNNs, Sherstinsky, 2020), a regularly spaced and gapless time dimension is usually required. Hence, data from an irregular time dimension should be aggregated or interpolated to fit into a regular temporal grid (e.g. gap-filling Landsat reflectances on a monthly basis, Moreno-Martínez et al., 2020). A suitable predefined temporal resolution must be selected, and data must be gap-filled. Various gap-filling techniques, ranging from simple linear interpolation to more complex AI-based modelling approaches, can be employed to address this (e.g. using Long Short-Term Memory networks, LSTMs, Ren et al., 2022). The choice of the gap-filling method depends on factors such as the data's nature, the desired accuracy level, and the specific requirements of the analysis or application.

3.5. Transformation

Expertly crafted higher-order features often prove highly relevant for addressing Earth system challenges. These new features span a spectrum, encompassing operations that range from simple transformations of the original variables to the creation of entirely novel features derived from advanced AI models. Examples of such features include the computation of spectral indices derived from reflectance bands (Montero et al., 2023), the extraction of frequencies through time series decomposition (Mahecha et al., 2010), the creation of spatio-temporal compositions (e.g. Griffiths et al., 2013), summarising high dimensional dynamics (e.g. Kraemer et al., 2020), and outputs generated by AI models (e.g. Brown et al., 2022). To illustrate, consider a study focusing on climate extremes like heatwaves and droughts' impact on the terrestrial biosphere. In such cases, calculating anomalies for critical variables (e.g. air temperature and soil moisture as proxies for heatwaves and droughts, with Gross Primary Production as the target biosphere variable) is pivotal (see Figure 3). Creating these novel features introduces a new dimension to distinguish between variable values corresponding to raw data and those representing anomalies. In line with the naming conventions introduced by Frantz, 2019, we designate an ESDC with higher-order features as a **hARDC Plus (hARDC+)**.

3.6. Reuse

ESDCs, after generation and analysis, can either evolve into dynamic versions through continuous updates or become static ESDCs, serving as input for the generation of new ESDCs. In the first scenario,

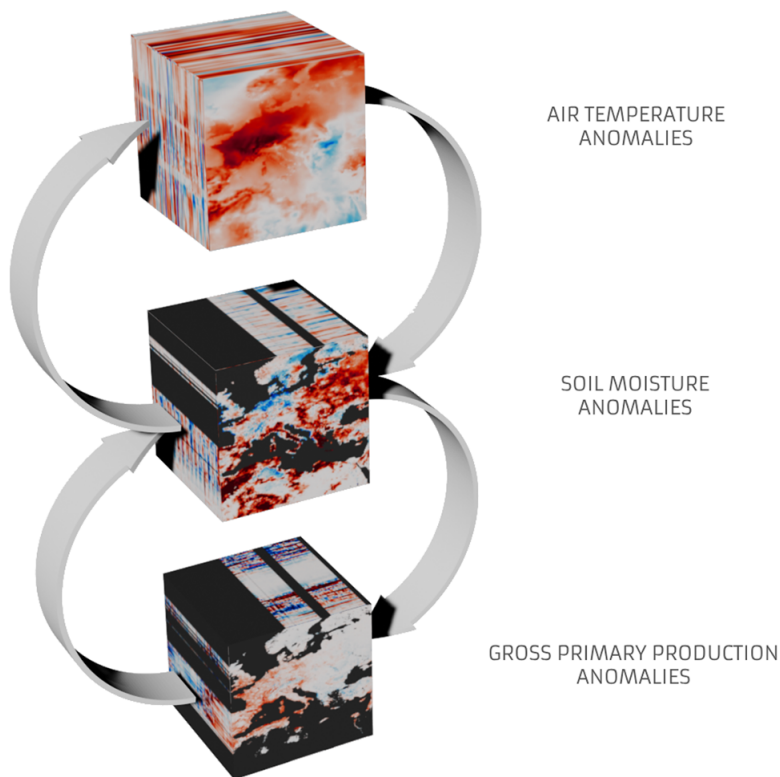


Figure 3. Abstract representation illustrating the connection between three Earth system variables in a hARDC+ (from top to bottom: anomalies in air temperature, soil moisture, and gross primary production). The arrows illustrate the interactions that can be modelled, e.g., predictive modelling (top to bottom) or interpretation (bottom to top), depending on the use case of interest.

establishing a Continuous Integration (CI) pipeline becomes essential for automating ESDC updates. This pipeline can be scheduled to align with the release of new dataset versions, ensuring the ESDC remains current. However, this approach may prove inefficient for EO products that are delivered regularly (e.g. Sentinel-2 or MODIS) and that may be constantly reprocessed by data providers, releasing new versions with updated processing pipelines. In this case, the update schedule should align with the specific needs of the ESDC (e.g. monthly, semi-annually, or annually). In the second scenario, an automatic update of dataset versions is also feasible, eliminating the necessity to extend the ESDC to the most recent date. In either scenario, it is crucial to implement clear reproducibility and traceability practices to ensure the data integrity of future ESDCs.

As highlighted in Section 3.1, standardisation is pivotal in promoting fluid data interoperability within this context. OGC has recognised the growing significance of data cube approaches for geospatial data. OGC recently established the GeoDataCubes SWG to create an API that facilitates interoperability among various solutions²⁶. This standard covers a broad scope, explicitly including API functionalities for access and processing, exchange format recommendations, profiles, and a metadata model.

Additionally, cloud technologies have ushered in the development of data cube services that abstract the underlying file structures and formats, replacing them with APIs offering diverse processing functionalities and promoting interoperability. For instance, platforms like Sentinel Hub²⁷ serve as sources for ESDC generation through tools like xcube. Moreover, the openEO platform²⁸ aims to provide an API that enables connections from multiple clients to various cloud backends using a unified API (Schramm et al., 2021). Approaches like these allow the tailored specification of ESDCs on-demand, with server-side processing relieving requesters of the complexities of the generation task. However, this convenience often comes with a trade-off, as the processing engine's code basis, the processing environment, and the input data are not known to requesters. Any modifications to these specifications can result in different outcomes for identical requests to the data cube API, hindering a streamlined update of dynamic ESDCs and a transparent basis for reusing static ESDCs.

In contrast, less convenient but more transparent approaches fully document the ESDC generation process through “recipes”. These recipes contain versioned source code used for input data processing. Examples include the Pangeo Forge²⁹ (Stern et al., 2022) and DeepESDL recipes³⁰. Recipes, coupled with versioned input data and fully specified processing environments, enable practical reproducibility of resulting ESDCs. This approach supports the seamless updating of dynamic ESDCs when new data becomes available and provides transparency for incorporating static ESDCs into new datasets.

Ongoing efforts to enhance data lineage and provenance transparency are integral to the Copernicus Data Space Ecosystem. The development of the “traceability” service³¹, currently in progress, is designed to empower users to trace all modifications to the data from its origin to its delivery to the end user, ensuring greater transparency and accountability in the ESDC life cycle.

3.7. Metadata generation

Traceability and self-explanatory power are essential aspects alongside the data values themselves. When an ESDC is generated, end users may access its description through various sources, including documentation that adheres to best practices for open data publishing within the Earth sciences. Such practices are supported by data journals (e.g. ESSD³²) and scientific associations (e.g. AGU Open Science³³), provided that the data producers have furnished comprehensive documentation. However, the data must

²⁶ <https://www.ogc.org/press-release/ogc-forms-new-geodatacube-standards-working-group>

²⁷ <https://www.sentinel-hub.com/>

²⁸ <https://openeo.cloud>

²⁹ <https://pangeo-forge.org>

³⁰ <https://github.com/deepesdl/cube-gen>

³¹ <https://dataspace.copernicus.eu/analyse/traceability>

³² https://www.earth-system-science-data.net/policies/data_policy.html

³³ <https://www.agu.org/-/media/files/publications/your-6-step-guide-for-publishing-open-access-with-agu.pdf/>

carry its own encapsulated description in the form of metadata, which typically comprises a set of attributes represented as key-value pairs. This ensures the data contain relevant information about their characteristics, facilitating understanding and utilisation.

Metadata generation should begin at the initial stage of data collection, encompassing crucial information such as data descriptors (e.g. name, units, measurement methods and equipment, resolution), data transformations (e.g. resampling or interpolation methods), metadata transformations (e.g. renaming procedures, conventions conversion), and responsible producers (e.g. creator entity, data provider). This metadata generation process should be consistently maintained throughout the entire ESDC life cycle, documenting each step undertaken to derive the final product (e.g. storing the process graphs from openEO when using this platform³⁴). This ensures comprehensive self-contained documentation of the history and processing of the ESDC.

While flexibility exists in metadata management, conventions are crucial when dealing with Earth system data. The Climate and Forecast Metadata Conventions (CF Conventions, Hassell et al., 2017), for instance, represent a comprehensive set of standards specifically designed for Earth system data stored in formats such as NetCDF (although they can be readily applied to other formats like Zarr). These conventions facilitate the creation of clear and detailed descriptions of data variables and coordinate dimensions. Furthermore, software like xarray (Hoyer and Hamman, 2017) can parse CF Conventions and leverage them for different ESDC processes³⁵. Compliance with CF Conventions not only simplifies data sharing but also promotes interoperability among various data sources, ensuring that ESDCs adhere to established standards.

4. Leveraging ESDCs for Earth system research

ESDCs offer promising opportunities for advancing Earth system research, particularly with recent AI developments. This is exemplified for Deep Learning (DL) by the spatio-temporal nature of ESDCs in a tensor-like structure. In this context, several key subjects emerge as highly relevant for Earth system research. We present three pertinent topics where the potential of ESDCs can be leveraged for advancing Earth system research: Physics-Informed Machine Learning (PIML), the adoption of complex sampling strategies, and the quantification of uncertainties.

4.1. Adding factual knowledge via PIML

A great addition to Machine Learning (ML) modelling is combining the pure data-driven approach with factual knowledge of the system under investigation (Karniadakis et al., 2021). PIML leverages domain knowledge (typically mechanistic models or differential equations) and flexible data-driven ML methods (typically neural networks). Consequently, PIML models respect physical boundaries more faithfully while being flexible enough to approximate arbitrarily complex non-linear functions from data (cf. discussion and references in Reichstein et al., 2019). ESDCs provide a unique structure to access multiple Earth system data streams, and the equation-based model describes the underlying process. Thanks to this ready availability of data and equations, exploring PIML models using a wide array of baseline models would be far easier and faster. The equations detailing a given variable could be added to the cube as a sub-field of the variable of interest in the same way that space and time are. The eventual implementation should consider the multi-platform and multi-language nature of ESDCs. As illustrated above, this requires a unified and robust approach that suits multiple use cases.

4.2. Sampling for AI in a complex system

Sampling on ESDCs is essential for learning the concrete interactions of drivers, spatial conditions, timing, and other determinants of specific processes and their implications. This involves strategically

³⁴ <https://api.openeo.org/v/0.3.0/processgraphs/>

³⁵ <https://docs.xarray.dev/en/stable/user-guide/weather-climate.html>

selecting a manageable subset from the ESDC. This selection process is particularly important for ML algorithms, as they rely on these subsets to establish a foundational understanding of the process to be analysed (Atkinson et al., 2022; Nikparvar and Thill, 2021). Pseudo-random sampling facilitates a broad and diverse data selection, while regionalised sampling uses specific patterns within the data for a more targeted analysis. The latter proves particularly advantageous when the research goal is to comprehend specific phenomena.

Constructing representative samples in Earth system processes must ensure an unbiased representation of the target variable. The multidimensional nature of Earth system processes poses sampling challenges across multiple variables. Consider, for instance, a study that aims at understanding the effects of climate extremes on the terrestrial biosphere using AI (Sippel et al., 2018). We know that climate extremes such as heatwaves, droughts, extreme precipitation, flooding, etc., are typically associated with multiple variables (Flach et al., 2021). Additionally, such events can co-occur in unfavourable sequences, i.e., compounding heatwaves, droughts, or floods following droughts (Zscheischler et al., 2020). To understand such circumstances, one should consider the full spatio-temporal extended in all relevant dimensions, including derived meta-variables that describe the characteristics of these events, such as timing, duration, extent, and intensity (Flach et al., 2017). Often, additional factors gain significance. For example, ecosystem responses to extremes vary in space depending on ecosystem conditions (Mahecha et al., 2017), land-cover types (Flach et al., 2021), and associated impacts, e.g., on the carbon cycle (Sippel et al., 2018). Building suitable AI models that predict such impacts requires including static data (e.g. vegetation type).

Yet, the critical question is then: how to obtain adequate and balanced training and validation data? Earth system processes often involve rare events of extreme conditions, which may occur sporadically over time and space. This rarity can lead to imbalanced datasets, where certain classes of the target variable or ranges of continuous values are underrepresented. This also applies to ranges of continuous values in an imbalanced distribution. Imbalanced datasets affect the performance and generalisation of models trained on these samples. Achieving spatio-temporal representativeness in this context can be challenging. To train ML algorithms for effective recognition and understanding of these events, it is crucial to include additional sampling within the specific domains where these events occur. For example, when constructing datasets for global flood (Li et al., 2023) or cloud detection (Aybar et al., 2022), the methodology involves initiating automatic sampling that covers a broad spectrum of ecosystem conditions. Simultaneously, manually selected events are introduced. This approach ensures a balanced representation of different classes in the dataset, thereby enhancing the algorithm's capability to accurately predict such events. Figure 4 showcases a potential workflow where event detection is performed based on global ESDCs, and samples for high-resolution ML are extracted based on a systematic sampling strategy (e.g. Ji et al., 2024). Here, analysing land cover purity is an option (a relatively homogeneous land cover dominated by a single vegetation type allows for easier comparisons and subsequent analyses), as well as incorporating mixed land covers (which introduces heterogeneity and interactions among land cover types), providing more comprehensive information for model training.

Finally, the selection of samples with the necessary data dimensions must align with the chosen algorithm. For instance, tabular-based algorithms like tree-based methods require 2-dimensional batches (sample and variable), which are selected as individual points from the spatio-temporal domain. DL methods like Transformers (Vaswani et al., 2017) or RNNs, e.g., LSTMs (Hochreiter and Schmidhuber, 1997), which consider sequence (or positional) dependencies, require 3-dimensional batches (e.g. sample, timestep, variable) and extract samples usually as subsets of time series from the spatial domain. Convolutional Neural Networks (CNNs, LeCun et al., 1989) may be used with 4-dimensional batches (e.g. sample, height, width, variable) by taking spatial subsets or grids from the temporal domain. DL methods accounting for both spatio-temporal dependencies, such as 3DCNNs or Convolutional LSTMs (ConvLSTMs, Shi et al., 2015), require 5-dimensional batches (e.g. sample, height, width, timestep, variable) and extract samples as subsets of ESDCs.

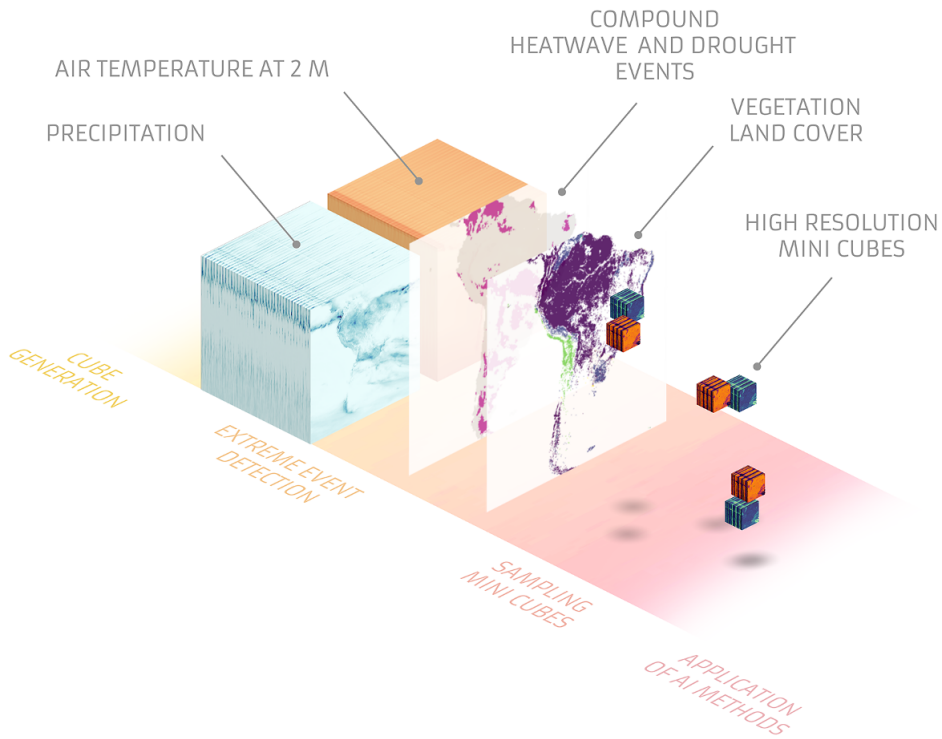


Figure 4. Abstract representation illustrating the process of sampling high-resolution mini cubes for further analysis by considering vegetation land covers and extreme events detected via a global ESDC. Note that sample mini cubes are specified in the spatial and temporal ranges of the detected extreme events (also considering their occurrence).

4.3. Quantifying uncertainties

Uncertainty quantification is crucial to Earth science, providing a comprehensive assessment of the reliability and confidence associated with scientific predictions, model simulations, and observational data. Capturing and modelling uncertainty is a complex task as it arises from various sources such as data limitations, model approximations, and the inherent complexity of Earth system dynamics.

Uncertainty can be broadly categorised into two types: epistemic uncertainty and aleatoric uncertainty (Kiureghian and Ditlevsen, 2009). Epistemic uncertainty refers to the model's confidence in its predictions and is related to the choice of model parameters. Techniques such as Bayesian inference or Dropout can estimate epistemic uncertainty (Srivastava et al., 2014; Gal and Ghahramani, 2016). Bayesian methods assign probability distributions to model parameters, directly quantifying uncertainty. In DL, dropout-based methods create model ensembles by randomly dropping out units during training, providing a measure of uncertainty based on the variability among the ensemble members. While these techniques may not completely capture the underlying uncertainty due to assumptions made during modelling or training, they are practical and can be employed to estimate uncertainty. These methods can be computationally demanding and time-consuming, mainly when applied to real-time applications. However, advancements in cloud platforms and the Monte Carlo (MC)-Dropout technique have enabled reliable uncertainty estimates, even when working with massive amounts of data (Martínez-Ferrer et al., 2022). On the other hand, aleatoric uncertainty is associated with the noise or variability present in the data (e.g. data affected by natural variability, measurement errors, or other sources of noise) and cannot be reduced. Instead, it can be identified and quantified as part of the uncertainty characterisation.

ESDCs involving measurements or modelled data can be accompanied by associated uncertainty values. Data assimilation techniques are key in incorporating data into ESDCs while considering the associated uncertainties. Approaches such as Kalman filtering, variational data assimilation, or ensemble-based assimilation can effectively merge different data sources and quantify the resulting uncertainties (Mathieu and O'Neill, 2008).

5. Challenges in ESDC analysis

While ESDCs present significant opportunities, it's crucial to approach them with a well-informed strategy to avoid naive applications of analytical methods. In this section, we describe challenges associated with ESDC analysis, focusing on two key issues: addressing geometric distortions (introduced during the cubing process) and spatio-temporal autocorrelation problems.

5.1. Geometric challenge on planet Earth

Most ESDCs covering the whole globe use a simple longitude-latitude plate-carrée projection, which fits the ESDC model very well. The approach also allows for efficient storage and subsetting of cubes to user-generated subsets corresponding to a bounding box. However, for advanced data analysis, equirectangular projections have two main drawbacks: 1) grid cells differing in latitude do not have equal area, and 2) the distances to nearest neighbours are not constant.

The first drawback introduces a sampling bias towards high latitudes in the data. This bias can affect the representativeness and accuracy of analyses (cf. Section 5.2), particularly for regions located closer to the equator. The most trivial cases are computations of scalars, like global means (e.g. Figure 5), which need to be weighted or approaches like principal component analyses that require area-weighted covariance matrices. Effects of this kind have been known for decades and are considered climate textbook knowledge (Storch et al., 2000). However, they remain a challenge, as we find them often ignored in ESDC analytics. Issues of this kind can be alleviated using area-weighted statistics, suitable for most linear algorithms, or by performing weighted sampling from grid cells. For advanced, often non-linear data science methods, considering the spherical geometry is much more challenging, and careful consideration is advised before naive applications are performed. Even when applying area-weighted statistics correctly, oversampled areas lead to unnecessary increases in storage requirements and computation time.

The second drawback is particularly significant when applying spatial convolutions or moving window operations. To address this, several approaches can be employed. One option is to use Spherical Harmonics for simple convolutions, providing a transformation that respects the spherical nature of the data (Wieczorek and Meschede, 2018). Spherical Harmonics can also be used as coordinate embeddings

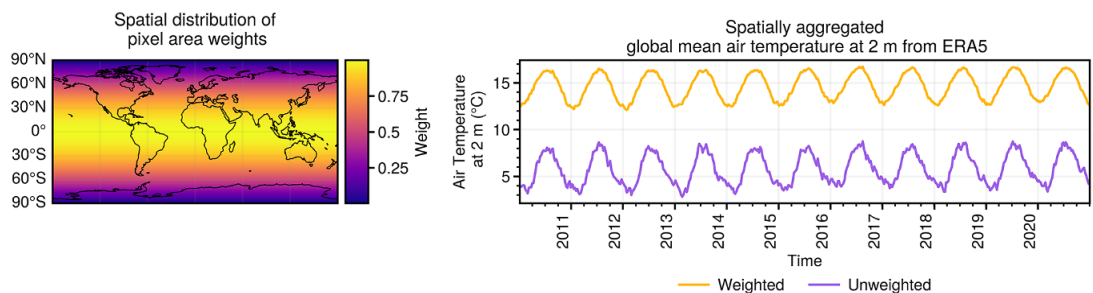


Figure 5. Comparison of air temperature at 2 m from ERA5 with and without weighting on the global mean time series computation. This rather trivial example shows how radically wrong any computation can be if the spherical nature of planet Earth is ignored.

for neural networks (Rußwurm et al., 2023). Another approach involves graph convolutions that consider varying distances to neighbours.

5.2. Spatio-temporal representativeness for an accurate model evaluation

Diagnostics on predictive modelling with ESDCs can be challenged by the representativeness and spatio-temporal structure of training data (Tobler, 1970; Meyer and Pebesma, 2021; Ploton et al., 2020; Kattenborn et al., 2022). Assessing the accuracy of a prediction is statistically straightforward as long as reference data is available for the entire population or if a respective sample represents the spatio-temporal structure of the population (Wadoux et al., 2021; Brus, 2021). However, many modelling tasks build on observations not representative of underlying temporal dynamics or an entire land surface variability (e.g. upscaling functional ecosystem properties from sparse and clustered FLUXNET sites). Such an imbalance in reference data may not necessarily lead to a bias in model coefficients (Pabon-Moreno et al., 2022). However, it may lead to inflated prediction accuracy estimates, given the commonly limited capacities of ML to extrapolate into the unknown, where the predictor-response relationship may depart (Ludwig et al., 2023). Thus, the accuracy assessment of a prediction estimated from clustered samples will not represent the factual accuracy of predictions beyond the reference data availability. This is critical for assessing the quality of a prediction itself and potential error propagation in subsequent analysis (Yates et al., 2018; Meyer and Pebesma, 2021; Mila et al., 2022). It is advised that predictions should inform on the area of applicability (Meyer and Pebesma, 2021), i.e., the area in which the predictor-space is covered by the reference data and obtained predictive accuracies thereof are assumed to hold.

However, assessing the predictive performance of a model inside the area of applicability may be challenged by the spatio-temporal structure of the training and test data. Commonly, adjacent observations (both in time and space) are more similar (autocorrelated in space and time), and therefore accuracies determined from test observations near the training data will be more accurate (Roberts et al., 2017; Dormann et al., 2007). For instance, seasonal effects can inflate model performance when using test observations near training data in the temporal dimension. Dependence among training and reference data results in any case on optimistic estimates of model performance, meaning that such accuracies do not reflect the actual transferability of the model to unseen areas or time steps (Roberts et al., 2017). For instance, Ploton et al. (2020) showed that ML-based models found accurate in the presence of spatial dependent training and validation data may learn spatial data structures instead of transferable relationships between a response (biomass) and the predictors (environmental variables and optical reflectance). This may not only lead to erroneous model transferability and extrapolation to new spatial or temporal domains but also prevent an adequate interpretation of model functioning and attribution to variables and processes (Sweet et al., 2023). Therefore, model performance and interpretation should be performed by minimising spatio-temporal dependence of observations via cross-validation strategies (cf. Roberts et al., 2017; Meyer et al., 2018; Ploton et al., 2020; Kattenborn et al., 2022).

6. Technical considerations for managing ESDCs

Managing ESDCs throughout their entire life cycle is complex and resource-intensive. This section outlines the technical considerations and limitations associated with the current state-of-the-art technological resources for ESDC management. This encompasses aspects such as computing resources, software tools, and scalable solutions that are crucial for effectively handling the challenges involved in ESDC management.

6.1. Computing resources

The data size and available computing resources determine data processing feasibility throughout the ESDC life cycle. Computing resources vary from a single laptop to a local cluster with multi-threaded or distributed processing capabilities and can extend to cloud computing environments composed of

multiple clusters. Modern computers are equipped with high-speed Solid-State Drives (SSDs) featuring fast random access and the potential for multiple Gigabytes per second throughput. However, the challenge lies in their limited capacity. In data centres, this is solved by using arrays of disks, but this introduces additional challenges, including latency, throughput, reliability, and security. Computation on local systems typically involves single-threaded or lightly multi-threaded computations with a higher level of interactivity. In High-Performance Computing (HPC) environments, the software operates in a multi-threaded or multi-core manner and is usually installed by a local system administrator. HPC environments are well-suited for extensive processing tasks but offer reduced interactivity due to the involvement of job schedulers for managing computation resources. Cloud computing environments offer a promising solution for managing vast amounts of Earth system data. These environments can be further improved in terms of scalability by utilising technologies like Kubernetes and Argo, which allow for specialised workflows. Platforms such as GEE, the European Open Science Cloud (EOSC)³⁶, Google Colaboratory³⁷, Amazon SageMaker³⁸, DeepESDL³⁹, Copernicus Data Space Ecosystem (CDSE)⁴⁰, and Kaggle⁴¹ provide opportunities for efficient data storage, processing, and collaboration in scientific research. However, it is essential to note that these platforms often impose certain limitations on the users. These limitations include storage capacity, computational resources, available tools for ESDC management, access permissions, and usage restrictions.

6.2. Software capabilities

In the context of managing ESDCs, diverse tools are available. Here, we present a compendium of useful tools for processing Earth system data within the ESDC life cycle in three prominent programming languages: Python, R, and Julia.

Python, arguably the most used language for ESDC management, offers `xarray` with labelled multidimensional arrays (Hoyer and Hamman, 2017), built on top of `numpy` (Harris et al., 2020), and supporting on-disk reading and parallel processing via `dask` (Rocklin, 2015) (a Python library for parallel computing, enhancing array objects by employing data partitioning into chunks and employing dynamic task scheduling). Multiple tools are tailored to construct and process `xarray` datasets, which represent ESDCs. For data collection, `rasterio` (Gillies et al., 2013), `rioxarray`⁴², `satpy` (Raspaud et al., 2023), or `EOREADER` (Maxant et al., 2022) are instrumental for reading GeoTIFFs and COGs, returning `xarray` objects. `Xarray` excels in reading NetCDF files and cloud-based data via `zarr` as `dask`-arrays. Vector data can be converted into `xarray` objects using `geocube` (Snow et al., 2023). Data sourced from STAC catalogues can be sought through `pystac-client` and directly transformed into `xarray` objects via `stackstac`⁴³, `odc-stac`⁴⁴, or `cubo` (Montero et al., 2024). These tools support data collection and immediate cubing, including the temporal dimension. GEE enables data retrieval as `numpy` arrays through its API, which can be directly converted into `xarray` objects using `Xee` or `wxee`. GEE's API (Gorelick et al., 2017) and extensions (Montero, 2021) allow data curation before cubing. `Xcube` has various data stores for data acquisition and `xarray` object generation⁴⁵. `XDGGS` (Kmoch et al., 2024) simplifies working with different DGGS in `xarray`. The curation, harmonisation, and transformation stages, being subjective and application-dependent, can be accomplished through `xarray` or `numpy` processing. Libraries like `scipy` (Virtanen et al., 2020), built

³⁶ <https://eosc-portal.eu/>

³⁷ <https://colab.research.google.com/>

³⁸ <https://aws.amazon.com/sagemaker/>

³⁹ <https://www.earthsystemdatalab.net/>

⁴⁰ <https://dataspace.copernicus.eu/>

⁴¹ <https://www.kaggle.com/>

⁴² <https://github.com/corteva/rioxarray>

⁴³ <https://github.com/gjoseph92/stackstac>

⁴⁴ <https://github.com/opendatacube/odc-stac>

⁴⁵ <https://xcube.readthedocs.io/en/latest/plugins.html>

on top of `numpy`, offer additional resources leveraging ESDCs as multidimensional arrays. The analysis phase leverages a plethora of tools. ESDCs as multidimensional arrays are compatible with `numpy`, `scipy`, and related tools. Moreover, ESDCs represented as tensors interface effectively with `tensorflow` (Abadi et al., 2016) or `pytorch` (Paszke et al., 2019). Furthermore, developments that aren't designed for direct ESDC use can also be leveraged using tensors in the representation of ESDCs (e.g. `torchgeo`, Stewart et al., 2022, `GeoTorchAI`, Chowdhury and Sarwat, 2022, `pytorch-metric-learning`, (Musgrave et al., 2020) and `TorchIO`, (Pérez-García et al., 2021)).

R, a widely used programming language for statistical analysis, has assumed increasing significance in geospatial data processing and management. Raster data sourced from image collections can be managed seamlessly, progressing from data collection to study, with the assistance of libraries like `raster`⁴⁶ or its more recent counterpart, `terra`⁴⁷. ESDCs can be collected and analysed through dedicated tools like `gdalcubes` (Appel and Pebesma, 2019) and `stars` (Pebesma and Bivand, 2023). Regionalised sampling using geospatial data can be conducted using `stpp` (Gabriel et al., 2013) and `spatstat` (Baddeley et al., 2015). Recent developments have introduced the capability for lazy on-disk reading of Zarr files⁴⁸. Furthermore, data can be sourced and cubed directly from STAC catalogues using `rstac` (Simoes et al., 2021b) in combination with `gdalcubes`. Another comprehensive package for ESDC management is `sits` (Simoes et al., 2021a), offering an end-to-end solution that additionally includes various tools for AI-related tasks, encompassing sampling, tuning, prediction, and the computation of uncertainty values.

Julia, a high-speed programming language, has gained popularity in scientific computing, making it an excellent choice for processing the large volumes of data found in ESDCs. Julia offers tools that cover crucial parts of the ESDC life cycle. These tools include `YAXArrays.jl`⁴⁹ and `Rasters.jl`⁵⁰ for multidimensional labelled array operations, `GriddingMachine.jl` (Wang et al., 2022) for data acquisition, and experimental libraries like `STAC.jl`⁵¹ for data discovery within STAC catalogues. For analysis, Julia provides specialised tools such as `EarthDataLab.jl`⁵² for the direct processing of the Earth System Data Cube (Mahecha et al., 2020). Moreover, data distortions introduced during the cubing process can be addressed using libraries like `OnlineStats.jl`⁵³ (Day and Zhou, 2020) and `WeightedOnlineStats.jl`⁵⁴ (Kraemer et al., 2020). Julia's ecosystem also includes ML tools like `Flux.jl` (Innes, 2018), `DiffEqFlux.jl` (Rackauckas et al., 2019), and `ReservoirComputing.jl` (Martinuzzi et al., 2022), enabling advanced data analysis, including novel techniques like PIML.

6.3. Scalability obstacles

The size of ESDCs poses several challenges for analysis. Generally, in most programming languages for data science (e.g. Python, Julia, R), data has to be completely loaded into memory before calculating a simple statistic (e.g. median). However, ESDCs often surpass the memory limit, hindering computations or resulting in significant slowdowns due to frequent disk read-write operations. Instead, users can apply specialised algorithms that calculate statistics iteratively (Welford, 1962; Schubert and Gertz, 2018). memory algorithms allow the user to track statistics (e.g. mean, sums, and standard deviations) iteratively. They give the user complete control (and responsibility) over the order of the data reads. Because of the spherical nature of the Earth and the resulting differences in the area covered by pixels, these computations

⁴⁶ [https://github.com/rspatial/raster](https://github.com/rsatial/raster)

⁴⁷ <https://github.com/rspatial/terra>

⁴⁸ <https://www.r-bloggers.com/2022/09/reading-zarr-files-with-r-package-stars/>

⁴⁹ <https://github.com/JuliaDataCubes/YAXArrays.jl>

⁵⁰ <https://github.com/rafaqz/Rasters.jl>

⁵¹ <https://github.com/JuliaClimate/STAC.jl>

⁵² <https://github.com/JuliaDataCubes/EarthDataLab.jl>

⁵³ <https://github.com/joshday/OnlineStats.jl>

⁵⁴ <https://github.com/gdkrmr/WeightedOnlineStats.jl>

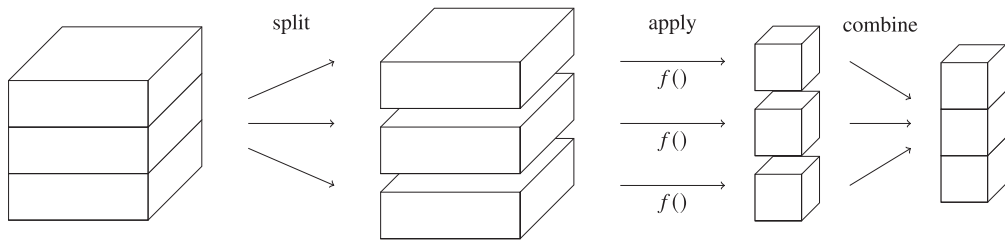


Figure 6. Split-apply-combine: split an ESDC along arbitrary axes, apply a function f to each sub-cube, and then combine the results along the same axes that have been used to split the original ESDC.

require weighted versions of the statistics (cf. Section 5.1). Errors arising from floating-point arithmetic must be minimised, including the potential for catastrophic cancellation (Kahan, 1965; Goldberg, 1991).

Often, analyses can be performed independently on timesteps, maps, or any other discrete chunks of an ESDC (e.g. dimensions, periods, spatial slices). First, users *split* the data into those chunks, and then *apply* the transformation. In the end, users *combine* the elements back together into a new ESDC (see Figure 6). Many analyses can be expressed in terms of *split-apply-combine* (Wickham, 2011; Mahecha et al., 2020), such as calculating mean seasonal cycle maps from a time axis to a day-of-year axis, or a global mean temperature time series that collapses latitude and longitude into a scalar value per timestep. This method is also known as *map-reduce* in distributed data processing. Still, in contrast, it is made for array-like or tabular data (and the *reduce* step always consists in concatenating the results of the *map* step, cf. Wickham, 2011). Implementations of *split-apply-combine* can trade-off between memory consumption and performance by adjusting the amount of data being loaded into memory simultaneously. They may also take advantage of parallel reading, processing, and writing of data, which is especially important if the data is not stored on local storage but on object stores with high access latency.

Storage in the form of compressed chunks typically employed by ESDCs, where reading a single element requires loading an entire chunk into memory, presents an opportunity for optimising sampling during ML training. Reading points individually is inefficient, as sampling two points from the same chunk necessitates reading the entire chunk twice. To mitigate this, reordering the points within a batch enables reading points from the same chunk jointly, reducing the number of reading operations. Adopting this approach makes it possible to limit the need to read the entire ESDC only once per batch, optimising the data access process.

Ensuring that scalability obstacles are transparent for end users during Earth system data analysis is essential. While experienced users may be able to address scalability issues effectively, less experienced users may struggle with the process if it is not fully transparent. It is important to provide a user-friendly interface that hides the complexities of scalability, allowing users to focus on their analysis tasks. Not all users can access sufficient computing resources for scaling processes, resulting in additional processing costs. Therefore, providing accessible and cost-effective solutions for scalability, such as cloud-based platforms, is crucial to enable a broader range of users to harness the benefits of scaling in Earth system data analysis.

7. Visual interaction with ESDCs

Data and process visualisation are critical for communicating Earth system science because big data are often hard to understand intuitively based on metadata alone, especially for non-expert audiences (Hibbard et al., 2002; Kendall et al., 2008; Kehrer and Hauser, 2012). The gap between analytic capability and the means to effectively visualise results slows our progress in understanding complex Earth system phenomena. Specialised tools are needed to visualise ESDCs and address their specific needs. Helbig et al. (2017) defined the key challenges of data visualisation for advancing Earth system sciences. Their ambition was to use ESDC visualisation for visual data exploration, facilitating multidisciplinary and collaborative research and also emphasising their educational role.

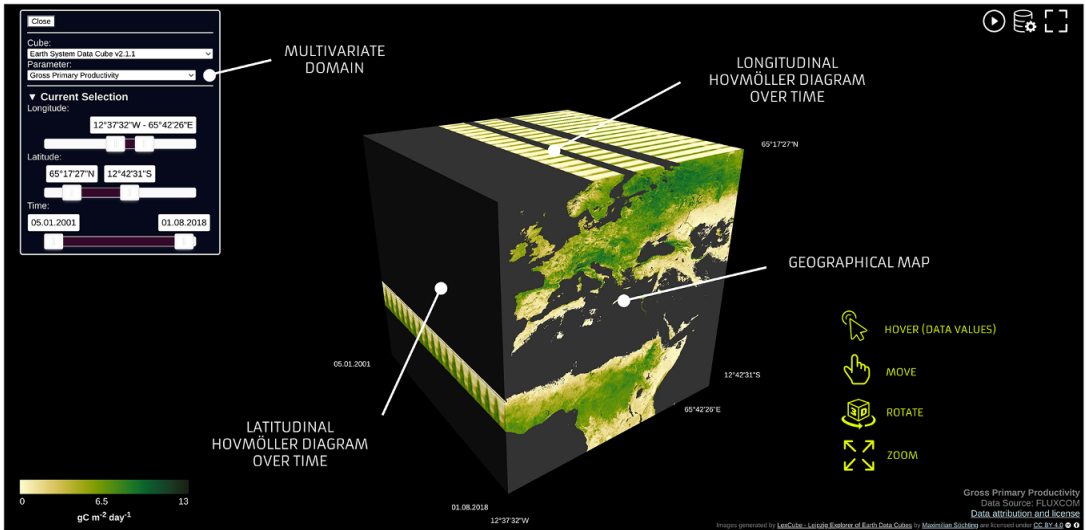


Figure 7. Interactions within an example ESDC in Lexcube, showcasing a geographical map on the front side and Hovmöller diagrams depicting temporal changes on the lateral sides. The ESDC allows for interactive subset operations on any side.

Much progress has been made in visualising ESDCs in Earth system research. Several viewers now have provided researchers with the means to explore and visualise multidimensional environmental datasets and generate scientific illustrations for publications^{55,56}. However, most approaches still rely on the classical geographical interpretation of georeferenced data and are restricted to displaying maps, extracting singular time series, or Hovmöller diagrams. Little advances have been made to visualise ESDCs, particularly multivariate ESDCs, for a better data understanding (cf. static attempts, Mahecha et al., 2010; Mahecha, 2017; Mahecha et al., 2020). The long-standing challenge is the trade-off between data interactions not designed for ESDCs and reliance on standard libraries that generate only static visualisations. Recent developments like Lexcube (Söchting et al., 2023, cf. Interactions in Figure 7)⁵⁷ and xcube-viewer⁵⁸ enable interactive and barrier-free visualisation, allowing users to inspect any ESDC dimension (especially space, time, and variable) interactively. Enabling interactions on large-scale spatio-temporal data on the web is key to democratising our science (Steed et al., 2014).

A significant challenge will be the integration of data analytics with interactive visualisations through visual analytics (cf. the review of Cui, 2019). The existing suite of methods is only partially suited for dealing with highly multivariate ESDCs, and most sophisticated visual analytic tools depend on a highly developed local computing infrastructure. There is a pressing need for web-based solutions to address this limitation. The goal should be to incorporate visualisations into any complex workflow to enhance comprehension of data inputs, monitor intermediate outcomes, and observe spatiotemporally structured results. One approach could be the tight integration of visualisation in developer workflows, particularly in popular environments like Jupyter Notebooks.

Integrating analytics tools with visualisation frameworks would allow researchers to dynamically explore, analyse, and visualise ESDCs in a unified environment in real-time. This would empower researchers to gain immediate insights into the relationships and patterns within the data. Additionally, incorporating visualisation into developer workflows would facilitate seamless visualisation generation at

⁵⁵ <https://github.com/carbonplan/maps>

⁵⁶ <https://cfs.climate.esa.int/>

⁵⁷ <https://www.lexcube.org/>

⁵⁸ <https://github.com/dcs4cop/xcube-viewer>

any stage of the ESDC life cycle, allowing researchers to visualise intermediate and final results and facilitating a more intuitive, iterative exploration of Earth system data.

ESDC visualisation extends its potential beyond the scientific community to engage and inform a wider audience. Nevertheless, this is particularly effective when accompanied by expert guidance such as tutorials, workshops, or annotations. Interactive open-access visualisations, exemplified by tools like Lexcube, allow political stakeholders and the general public to directly access and examine climate data (e.g. global or regional climate anomalies and trends). Open-access interactive visualisations enable scientifically literate individuals and those with less technical expertise to delve into ESDCs easily and rapidly by visualising anomalies, trends, and the interplay of variables. Such accessibility encourages a broader understanding and appreciation of Earth system research among diverse stakeholders, fostering a more informed and constructive dialogue about climate-related issues.

8. Conclusions and perspective

This paper reviews and explores the challenges and opportunities of leveraging ESDCs for Earth system research. This becomes particularly important in developing Earth Digital Twins (i.e. “a digital replication of the state and temporal evolution of the Earth system”, Bauer et al., 2021b). In this sense, the topics discussed here are of significance in initiatives like Destination Earth (DestinE)⁵⁹. The inherent simplicity and versatility of ESDCs enable a comprehensive exploration of the complex Earth system, facilitating a deeper understanding of intricate processes and phenomena. For advancing our understanding of the Earth system, the following key considerations emerge and need to be addressed by the research community to tap into the full potential of ESDCs:

1. **Artificial Intelligence on ESDCs:** The abundance of large-scale Earth system data, coupled with recent advancements in AI methods, compels the application of the latest developments in deep learning to ESDCs. Capitalising on the tensor-like structure of ESDCs in DL and incorporating factual knowledge through Physics-Informed Machine Learning approaches promise great advances in modelling and understanding. Recent advancements in AI, particularly in attention mechanisms, have opened up new possibilities for Earth system research. Techniques such as LLMs, generative image models (e.g. Stable Diffusion, Rombach et al., 2021), as well as recent image and video segmentation models (e.g. Segment Anything Model, SAM and SAM 2, Kirillov et al., 2023; Ravi et al., 2024), may hold the potential to significantly advance our understanding of the Earth system (Wu and Osco, 2023; Osco et al., 2023). The ability to ‘communicate’ to ESDCs to extract valuable insights (e.g. Lobry et al., 2020) is within reach (e.g. using text prompts to extract variable anomalies from a specific land cover over a particular region). Furthermore, there is potential to generate ESDCs using text prompts, images, videos, or additional data inputs simultaneously by leveraging the power of multi-modal mechanisms (e.g. ImageBind, Girdhar et al., 2023), e.g., simulating the impact on vegetation due to an extreme event over a real ESDC using text prompts and geographical data. However, caution must be exercised when applying AI methods to ESDCs to avoid erroneous predictions and interpretations. Factors such as spatio-temporal auto-correlation, the spherical nature of the Earth, and biased sampling in the spatio-temporal and multivariate domains pose risks. Still, the abstract nature of ESDCs provides an opportunity to establish a de facto standard for AI in Earth system science, benefiting from optimised data access and technical enhancements. To ensure reliable outcomes, standardised methods are needed to address spatial dependency, the model’s area of applicability, and model uncertainty within ESDC structures.
2. **Interacting with ESDCs:** The heterogeneity, size, and multivariate nature of datasets also may imply that using ESDCs’ is unintuitive, which hampers interpretation. Effective communication

⁵⁹ <https://digital-strategy.ec.europa.eu/en/policies/destination-earth>

opportunities with such data are crucial throughout the ESDC life cycle, both for scientists and a wider audience. Visualisation plays a key role in this regard. While visualisation tools are available to support the analysis process and scientific dissemination, there is still considerable potential for further exploration and development of visualisations. We believe that interactive visualisations are one key, as demonstrated by Lexcube. One promising avenue is the integration of visualisation directly into the analytics workflow (e.g. within Jupyter Notebooks or similar environments), and another is enabling visual analytics of ESDCs. In both cases, the challenge is making such interactions possible during the analysis process to enable the scientific exploitation of large ESDCs.

3. **Technical challenges of large ESDCs:** The multidimensional nature, varying spatio-temporal scales and resolutions, and applicability of ESDCs imply a series of technical challenges. These include interoperability issues, different geographical projections, interpolation and aggregation questions, and varying readiness levels for further analyses. Ensuring data integrity and interpretability while making Earth system data analysis-ready and interoperable requires tracing and encoding all data transformations and modifications in ESDC metadata. To address these challenges, developing guidelines and standards for geospatial datacubes is crucial for promoting FAIR and Open Earth System Science. The ever-increasing size and complexity of datasets demand scalable solutions to tackle associated challenges. The ongoing efforts of the open-source software community are commendable in this regard, as they contribute to the advancement of tools and frameworks tailored to handle big Earth system data. Furthermore, cloud environments present a possible solution to quickly scale workloads when processing data within the ESDC life cycle. They offer the advantages of on-demand resource allocation and scalability, allowing researchers to access the necessary computational power and storage capacity when needed.
4. **Integrating (geospatial) data beyond cubes:** ESDCs already offer the potential for advancing Earth system research and analysis in multiple domains. However, ESDCs can benefit from integrating different methodological approaches or data sources at different scales. One example is the integration of Unoccupied Aerial Vehicle (UAV)- and Light Detection and Ranging (LiDAR)-based data. This data provides a means to collect highly localised and high-resolution measurements, making them particularly suitable for localised studies and gaining valuable insights into fine-scale processes. Another example is the integration of vector data⁶⁰, which typically represents categorical information and carries great importance in multiple Earth system spheres (e.g. socioeconomic features). Additionally, in-situ collections of any process (e.g. via ecological monitoring data) are essential. Today, the quest is that users request the integration of any additional data sources while remaining fully valid. Yet, it poses a challenge as it raises important questions regarding interoperability and the encapsulation of multi-resolution cubes that incorporate multi-scale raster data and the combination of raster and vector data within a unified framework.
5. **Towards flexible cube-based structures:** To advance ESDCs' benefits, it is essential to advance the standards of ESDC structures and start considering hierarchical data structures, including ESDCs as "leaves" (e.g. xarray's DataTree structure) or even unstructured grid systems (e.g. Project Raijin⁶¹ with uxarray⁶²). Given the abundance of insightful (but heterogeneous) datasets, this would enhance Earth system research, regardless of their resolution or dimensionality. Nevertheless, this implies that we must ensure data traceability and interpretability as heterogeneity increases in the resolution or dimensionality domains. A prime example lies in integrating AI models' predictions within ESDCs. In such instances, additional dimensions must be incorporated to capture uncertainties (or quality flag systems) associated with AI-based predictions. This provides valuable insights into the reliability and robustness of the data. Leveraging the power of ESDCs in diverse fields can drive innovation, advance scientific knowledge, and enable more informed decision-making in a wide range of domains.

⁶⁰ <https://r-spatial.org/t/2022/09/12/vdc.html>

⁶¹ <https://raijin.ucar.edu/>

⁶² <https://github.com/UXARRAY/uxarray>

Open peer review. To view the open peer review materials for this article, please visit <http://doi.org/10.1017/eds.2024.22>.

Acknowledgments. We are grateful for the European Space Agency (ESA) funding for the DeepESDL and the DeepExtremes projects. Also, we thank the DLR for funding the ML4Earth and VW for funding the Digital Forest project. We also thank the DFG for supporting NFDI4Earth and NFDI4Biodiversity. We thank Pablo Mahecha for generating Figure 3 using inputs from Lexcube. Comments by the editor and the anonymous reviewers greatly improved the quality of the paper. Furthermore, we thank Peter Zellner for his comments and suggestions.

Author contribution. Conceptualisation: D.M.; M.D.M.; F.G.; G.K. Writing - Original Draft: D.M. with contributions from M.D.M.; G.K.; A.A.; C.A.; F.C.; I.F.; F.G.; S.H.; C.J.; T.K.; L.M.F.; F.M.; M.R.; M.S.; K.T. Review and Editing: D.M.; M.D.M.; G.K.; G.C.V.; T.K.; S.H. Visualisation: D.M.; M.D.M.; G.K.; C.J.; M.S. Supervision: M.D.M. All authors approved the final submitted draft.

Competing interest. The authors declare no competing interests exist.

Data availability statement. No data were used in this paper.

Ethical standard. The research meets all ethical guidelines, including adherence to the legal requirements of the study country.

Funding statement. This research was supported by grants from the European Space Agency ESA (“AI4Science - Deep Extremes” and “DeepESDL”). D.M. and M.D.M. acknowledge support from the “Digital Forest” project, Ministry of Lower-Saxony for Science and Culture (MWK) via the program Niedersächsisches Vorab (ZN 3679), and the “RS4BEF” project via the iDiv’s Flexpool program. M.D.M. and M.R. acknowledges support by the German Aerospace Center, DLR representing the Bundesministerium für Wirtschaft und Klimaschutz (ML4Earth, 50EE2201B). M.D.M., M.S., F.C. and F.G. acknowledge support by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) for funding the “NFDI4Earth”, project number: 460036893. T.K. acknowledges support by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) for funding the “PANOPS”, project number: 504978936. C.A. acknowledges support by the National Council of Science, Technology, and Technological Innovation (CONCYTEC, Peru) through the “PROYECTOS DE INVESTIGACIÓN BÁSICA – 2023-01” program with contract number PE501083135–2023-PROCIENCIA. M.D.M. and F.M. acknowledge support by the Federal Ministry of Education and Research of Germany and by Sächsische Staatsministerium für Wissenschaft, Kultur und Tourismus in the programme Center of Excellence for AI-research “Center for Scalable Data Analytics and Artificial Intelligence Dresden/Leipzig”, project identification number: ScaDS.AI.

References

- Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mane, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viegas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems.
- Appel, M. and Pebesma, E. (2019). On-demand processing of data cubes from satellite image collections with the gdalclouds library. *Data*, 4(3):92.
- Ariza-Porras, C., Bravo, G., Villamizar, M., Moreno, A., Castro, H., Galindo, G., Cabera, E., Valbuena, S., and Lozano, P. (2017). CDCol: A Geoscience Data Cube that Meets Colombian Needs. In *Communications in Computer and Information Science*, pages 87–99. Springer International Publishing.
- Atkinson, P. M., Stein, A., and Jeganathan, C. (2022). Spatial sampling, data models, spatial scale and ontologies: Interpreting spatial statistics and machine learning applied to satellite optical remote sensing. *Spatial Statistics*, 50:100646.
- Aybar, C., Ysuhuaylas, L., Loja, J., Gonzales, K., Herrera, F., Bautista, L., Yali, R., Flores, A., Diaz, L., Cuenca, N., Espinoza, W., Prudencio, F., Llactayo, V., Montero, D., Sudmanns, M., Tiede, D., Mateo-García, G., and Gómez-Chova, L. (2022). Cloudsen12, a global dataset for semantic understanding of cloud and cloud shadow in sentinel-2. *Scientific Data*, 9(1).
- Baddeley, A., Rubak, E., and Turner, R. (2015). *Spatial Point Patterns*. Chapman & Hall/CRC Interdisciplinary Statistics. Apple Academic Press, Oakville, MO.
- Bauer, P., Dueben, P. D., Hoefler, T., Quintino, T., Schulthess, T. C., and Wedi, N. P. (2021a). The digital revolution of earth-system science. *Nature Computational Science*, 1(2):104–113.
- Bauer, P., Stevens, B., and Hazeleger, W. (2021b). A digital twin of earth for the green transition. *Nature Climate Change*, 11(2): 80–83.
- Baumann, P., Dehmel, A., Furtado, P., Ritsch, R., and Widmann, N. (1998). The multidimensional database system RasDaMan. *ACM SIGMOD Record*, 27(2):575–577.
- Baumann, P., Misev, D., Merticariu, V., and Huu, B. P. (2019). Datacubes: Towards space/time analysis-ready data. *Service-Oriented Mapping: Changing Paradigm in Map Production and Geoinformation Management*, pages 269–299.

- Béjar, R., Lacasta, J., Lopez-Pellicer, F. J., and Noguerras-Iso, J. (2023). Discrete global grid systems with quadrangular cells as reference frameworks for the current generation of earth observation data cubes. *Environmental Modelling & Software*, 162: 105656.
- Boulton, G. (2018). The challenges of a big data earth. *Big Earth Data*, 2(1):1–7.
- Brandt, G., Balfanz, A., Fomferra, N., Harish, T. M., Mahecha, M., Kraemer, G., Montero, D., Meißl, S., Achtsnit, S., Umlauf, J., Neumann, A., Horton, A., Ewart, M., Gans, F., and Anghelea, A. (2023). DeepESDL – an open platform for research and collaboration in earth sciences.
- Bravo, G., Castro, H., Moreno, A., Ariza-Porras, C., Galindo, G., Cabrera, E., Valbuena, S., and Lozano-Rivera, P. (2017). Architecture for a colombian data cube using satellite imagery for environmental applications. In *Communications in Computer and Information Science*, pages 227–241. Springer International Publishing.
- Brown, C. F., Brumby, S. P., Guzder-Williams, B., Birch, T., Hyde, S. B., Mazzariello, J., Czerwinski, W., Pasquarella, V. J., Haertel, R., Ilyushchenko, S., Schwehr, K., Weisse, M., Stolle, F., Hanson, C., Guinan, O., Moore, R., and Tait, A. M. (2022). Dynamic World, Near real-time global 10m land use land cover mapping. *Scientific Data*, 9(1).
- Brus, D. J. (2021). Statistical approaches for spatial sample survey: Persistent misconceptions and new developments. *European Journal of Soil Science*, 72(2):686–703.
- Bueso, D., Piles, M., and Camps-Valls, G. (2020). Nonlinear pca for spatio-temporal analysis of earth observation data. *IEEE Transactions on Geoscience and Remote Sensing*, 58(8):5752–5763.
- Camps-Valls, G., Gerhardus, A., Ninad, U., Varando, G., Martius, G., Balaguer-Ballester, E., Vinuesa, R., Diaz, E., Zanna, L., and Runge, J. (2023). Discovering causal relations and equations from data. *arXiv preprint arXiv:2305.13341*.
- Camps-Valls, G., Svendsen, D. H., Cortes-Andres, J., Marenó-Martínez, A., Pérez-Suay, A., Adsuara, J., Martín, I., Piles, M., Muñoz-Mari, J., and Martino, L. (2021). Physics-Aware Machine Learning for Geosciences and Remote Sensing. In *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*. Ieee.
- Chadhuri, S. and Dayal, U. (1997). An overview of data warehousing and OLAP technology. *ACM SIGMOD Record*, 26(1): 65–74.
- Chen, X., Huang, Y., Nie, C., Zhang, S., Wang, G., Chen, S., and Chen, Z. (2022). A long-term reconstructed TROPOMI solar-induced fluorescence dataset using machine learning algorithms. *Scientific Data*, 9(1).
- Chowdhury, K. and Sarwat, M. (2022). Geotorch: A spatiotemporal deep learning framework. In *Proceedings of the 30th International Conference on Advances in Geographic Information Systems, SIGSPATIAL '22*. Association for Computing Machinery.
- Christiansen, R., Baumann, M., Kuemmerle, T., Mahecha, M. D., and Peters, J. (2022). Toward causal inference for spatio-temporal data: conflict and forest loss in colombia. *Journal of the American Statistical Association*, 117(538):591–601.
- Claverie, M., Ju, J., Masek, J. G., Dungan, J. L., Vermote, E. F., Roger, J.-C., Skakun, S. V., and Justice, C. (2018). The Harmonized Landsat and Sentinel-2 surface reflectance data set. *Remote Sensing of Environment*, 219:145–161.
- Clinton, N. (2023). Pixels to the people!
- Cortés-Andrés, J., Camps-Valls, G., Sippel, S., Székely, E., Sejdinovic, D., Diaz, E., Pérez-Suay, A., Li, Z., Mahecha, M., and Reichstein, M. (2022). Physics-aware nonparametric regression models for earth data analysis. *Environmental Research Letters*, 17(5):054034.
- Cracknell, A. P. (1998). Review article Synergy in remote sensing-whats in a pixel? *International Journal of Remote Sensing*, 19(11):2025–2047.
- Cui, W. (2019). Visual analytics: A comprehensive overview. *IEEE access*, 7:81555–81573.
- Day, J. and Zhou, H. (2020). OnlineStats.JI: A Julia Package for Statistics on Data Streams. *Journal of Open Source Software*, 5(46):1816.
- Dhu, T., Dunn, B., Lewis, B., Lymburner, L., Mueller, N., Telfer, E., Lewis, A., McIntyre, A., Minchin, S., and Phillips, C. (2017). Digital earth Australia – unlocking new value from earth observation data. *Big Earth Data*, 1(1–2):64–74.
- Dhu, T., Giuliani, G., Juárez, J., Kavvada, A., Killough, B., Merodio, P., Minchin, S., and Ramage, S. (2019). National Open Data Cubes and Their Contribution to Country-Level Development Policies and Practices. *Data*, 4(4):144.
- Dormann, C., M. McPherson, J., B. Araújo, M., Bivand, R., Bolliger, J., Carl, G., G. Davies, R., Hirzel, A., Jetz, W., Daniel Kissling, W., et al. (2007). Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography*, 30(5):609–628.
- Estupiñan Suarez, L. M., Gans, F., Brenning, A., Gutierrez-Velez, V. H., Londono, M. C., Pabon-Moreno, D. E., Poveda, G., Reichstein, M., Reu, B., Sierra, C. A., et al. (2021). A regional Earth system data lab for understanding ecosystem dynamics: An example from tropical South America. *Frontiers in Earth Science*, page 574.
- Estupinan-Suarez, L. M., Mahecha, M. D., Brenning, A., Kraemer, G., Poveda, G., Reichstein, M., and Sierra, C. A. (2023). Spatial patterns of vegetation activity related to enso in northern south america. *Journal of Geophysical Research: Biogeosciences*. In press.
- Ferreira, K. R., Queiroz, G. R., Vinhas, L., Marujo, R. F. B., Simoes, R. E. O., Picoli, M. C. A., Camara, G., Cartaxo, R., Gomes, V. C. F., Santos, L. A., Sanchez, A. H., Arcanjo, J. S., Fronza, J. G., Noronha, C. A., Costa, R. W., Zaglia, M. C., Zioti, F., Korting, T. S., Soares, A. R., Chaves, M. E. D., and Fonseca, L. M. G. (2020). Earth observation data cubes for brazil: Requirements, methodology and products. *Remote Sensing*, 12(24):4033.
- Flach, M., Brenning, A., Gans, F., Reichstein, M., Sippel, S., and Mahecha, M. D. (2021). Vegetation modulates the impact of climate extremes on gross primary production. *Biogeosciences*, 18(1):39–53.

- Flach, M., Gans, F., Brenning, A., Denzler, J., Reichstein, M., Rodner, E., Bathiany, S., Bodesheim, P., Guanche, Y., Sippel, S., et al. (2017). Multivariate anomaly detection for earth observations: a comparison of algorithms and feature extraction techniques. *Earth System Dynamics*, 8(3):677–696.
- Frantz, D. (2019). FORCE—Landsat + Sentinel-2 Analysis Ready Data and Beyond. *Remote Sensing*, 11(9).
- Gabriel, E., Rowlingson, B. S., and Diggle, P. J. (2013). stpp: An r package for plotting, simulating and analyzing spatio-temporal point patterns. *Journal of Statistical Software*, 53(2):1–29.
- Gal, Y. and Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. Proceedings of Machine Learning Research.
- Gillies, S. et al. (2013). Rasterio: geospatial raster i/o for Python programmers.
- Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K. V., Joulín, A., and Misra, I. (2023). Imagebind: One embedding space to bind them all.
- Giuliani, G., Chatenoux, B., Bono, A. D., Rodila, D., Richard, J.-P., Allenbach, K., Dao, H., and Peduzzi, P. (2017). Building an Earth Observations Data Cube: lessons learned from the Swiss Data Cube (SDC) on generating Analysis Ready Data (ARD). *Big Earth Data*, 1(1–2):100–117.
- Giuliani, G., Masó, J., Mazzetti, P., Nativi, S., and Zabala, A. (2019). Paving the way to increased interoperability of earth observations data cubes. *Data*, 4(3):113.
- Goldberg, D. (1991). What Every Computer Scientist Should Know about Floating-Point Arithmetic. *ACM Computing Surveys*, 23(1):5–48.
- Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., and Moore, R. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, 202:18–27.
- Griffiths, P., van der Linden, S., Kuemmerle, T., and Hostert, P. (2013). A Pixel-Based Landsat Compositing Algorithm for Large Area Land Cover Mapping. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 6(5):2088–2101.
- Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Ro, J. F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., and Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, 585(7825):357–362.
- Hassell, D., Gregory, J., Blower, J., Lawrence, B. N., and Taylor, K. E. (2017). A data model of the Climate and Forecast metadata conventions (CF-1.6) with a software implementation (cf-python v2.1). *Geoscientific Model Development*, 10(12):4619–4646.
- Helbig, C., Dransch, D., Böttinger, M., Devey, C., Haas, A., Hlawitschka, M., Kuenzer, C., Rink, K., Schäfer-Neth, C., Scheuermann, G., et al. (2017). Challenges and strategies for the visual exploration of complex environmental data. *International Journal of Digital Earth*, 10(10):1070–1076.
- Hibbard, B., Böttinger, M., Schultz, M., and Biercamp, J. (2002). Visualization in earth system science. *Acm Siggraph Computer Graphics*, 36(4):5–9.
- Hochreiter, S. and Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.
- Hoyer, S. and Hamman, J. (2017). xarray: N-d labeled arrays and datasets in python. *Journal of Open Research Software*, 5(1):10.
- Hsieh, W. W. (2022). Evolution of machine learning in environmental science—A perspective. *Environmental Data Science*, 1.
- Ilie, I., Dittrich, P., Carvalhais, N., Jung, M., Heinemeyer, A., Migliavacca, M., Morison, J. I., Sippel, S., Subke, J.-A., Wilkinson, M., et al. (2017). Reverse engineering model structures for soil and ecosystem respiration: the potential of gene expression programming. *Geoscientific Model Development*, 10(9):3519–3545.
- Innes, M. (2018). Flux: Elegant machine learning with julia. *Journal of Open Source Software*, 3(25):602.
- Irrgang, C., Boers, N., Sonnewald, M., Barnes, E. A., Kadow, C., Staneva, J., and Saynisch-Wagner, J. (2021). Towards neural earth system modelling by integrating artificial intelligence in earth system science. *Nature Machine Intelligence*, 3(8):667–674.
- Ji, C., Fincke, T., Benson, V., Camps-Valls, G., Fernandez-Torres, M.-A., Gans, F., Kraemer, G., Martinuzzi, F., Montero, D., Mora, K., Pellicer-Valero, O. J., Robin, C., Soechting, M., Weynants, M., and Mahecha, M. D. (2024). Deepextremecubes: Integrating earth system spatio-temporal data for impact assessment of climate extremes.
- Kahan, W. (1965). Further Remarks on Reducing Truncation Errors. *Communications of the ACM*, 8(1):40.
- Karniadakis, G. E., Kevrekidis, I. G., Lu, L., Perdikaris, P., Wang, S., and Yang, L. (2021). Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440.
- Kasprzyk, J.-P. and Donnay, J.-P. (2017). A raster solap designed for the emergency services of brussels agglomeration. In *CLOUD COMPUTING 2017 - The Eighth International Conference on Cloud Computing, GRIDS, and Virtualization*.
- Kattenborn, T., Schiefer, F., Frey, J., Feilhauer, H., Mahecha, M. D., and Dormann, C. F. (2022). Spatially autocorrelated training and validation samples inflate performance assessment of convolutional neural networks. *ISPRS Open Journal of Photogrammetry and Remote Sensing*, 5:100018.
- Kehrer, J. and Hauser, H. (2012). Visualization and visual analysis of multifaceted scientific data: A survey. *IEEE transactions on visualization and computer graphics*, 19(3):495–513.
- Kendall, W., Glatter, M., Huang, J., Hoffman, F., and Bernholdt, D. E. (2008). Web enabled collaborative climate visualization in the earth system grid. In *2008 International Symposium on Collaborative Technologies and Systems*, pages 212–220. IEEE.
- Killough, B. (2018). Overview of the Open Data Cube Initiative. In *IGARSS 2018–2018 IEEE International Geoscience and Remote Sensing Symposium*. Ieee.

- Killough, B.** (2019). The impact of analysis ready data in the africa regional data cube. In *IGARSS 2019–2019 IEEE International Geoscience and Remote Sensing Symposium*. IEEE.
- Killough, B., Siqueira, A., and Dyke, G.** (2020). Advancements in the open data cube and analysis ready data — past, present and future. In *IGARSS 2020–2020 IEEE International Geoscience and Remote Sensing Symposium*. IEEE.
- Kirilov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., and Girshick, R.** (2023). Segment anything.
- Kiureghian, A. D. and Ditlevsen, O.** (2009). Aleatory or epistemic? does it matter? *Structural Safety*, 31(2):105–112. Risk Acceptance and Risk Communication.
- Kmoch, A., Bovy, B., Magin, J., Abernathy, R., Coca-Castro, A., Strobl, P., Fouilloux, A., Loos, D., Uemaa, E., Chan, W. T., Delouis, J.-M., and Odaka, T.** (2024). Xdgs: A community-developed xarray package to support planetary dggs data cube computations. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLVIII-4/W12–2024:75–80.
- Kmoch, A., Vasilyev, I., Virro, H., and Uemaa, E.** (2022). Area and shape distortions in open-source discrete global grid systems. *Big Earth Data*, 6(3):256–275.
- Kopp, S., Becker, P., Doshi, A., Wright, D. J., Zhang, K., and Xu, H.** (2019). Achieving the full vision of earth observation data cubes. *Data*, 4(3):94.
- Kraemer, G., Camps-Valls, G., Reichstein, M., and Mahecha, M. D.** (2020). Summarizing the State of the Terrestrial Biosphere in Few Dimensions. *Biogeosciences*, 17(9):2397–2424.
- Krich, C., Migliavacca, M., Miralles, D. G., Kraemer, G., El-Madany, T. S., Reichstein, M., Runge, J., and Mahecha, M. D.** (2021). Functional convergence of biosphere–atmosphere interactions in response to meteorological conditions. *Biogeosciences*, 18(7):2379–2404.
- LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D.** (1989). Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation*, 1(4):541–551.
- Lewis, A., Oliver, S., Lymburner, L., Evans, B., Wyborn, L., Mueller, N., Raevksi, G., Hooke, J., Woodcock, R., Sixsmith, J., Wu, W., Tan, P., Li, F., Killough, B., Minchin, S., Roberts, D., Ayers, D., Bala, B., Dwyer, J., Dekker, A., Dhu, T., Hicks, A., Ip, A., Purss, M., Richards, C., Sagar, S., Trenham, C., Wang, P., and Wang, L.-W.** (2017). The Australian Geoscience Data Cube — Foundations and lessons learned. *Remote Sensing of Environment*, 202:276–292.
- Li, X. and Xiao, J.** (2019). A global, 0.05-degree product of solar-induced chlorophyll fluorescence derived from OCO-2, MODIS, and reanalysis data. *Remote Sensing*, 11(5):517.
- Li, Y., Dang, B., Li, W., and Zhang, Y.** (2023). Glh-water: A large-scale dataset for global surface water detection in large-size very-high-resolution satellite imagery.
- Lobry, S., Marcos, D., Murray, J., and Tuia, D.** (2020). Rsvqa: Visual question answering for remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 58(12):8555–8566.
- Ludwig, M., Moreno-Martinez, A., Hölzel, N., Pebesma, E., and Meyer, H.** (2023). Assessing and improving the transferability of current global spatial prediction models. *Global Ecology and Biogeography*, 32(3):356–368.
- Mahecha, M.** (2017). Earth system data cube. figshare. Figure. DOI: 10.6084/m9.figshare.4822930.v2.
- Mahecha, M. D., Fürst, L. M., Gobron, N., and Lange, H.** (2010). Identifying multiple spatiotemporal patterns: A refined view on terrestrial photosynthetic activity. *Pattern Recognition Letters*, 31(14):2309–2317.
- Mahecha, M. D., Gans, F., Brandt, G., Christiansen, R., Cornell, S. E., Fomferra, N., Kraemer, G., Peters, J., Bodesheim, P., Camps-Valls, G., et al.** (2020). Earth system data cubes unravel global multivariate dynamics. *Earth System Dynamics*, 11(1):201–234.
- Mahecha, M. D., Gans, F., Sippel, S., Donges, J. F., Kaminski, T., Metzger, S., Migliavacca, M., Papale, D., Rammig, A., and Zscheischler, J.** (2017). Detecting impacts of extreme events with ecological in situ monitoring networks. *Biogeosciences*, 14(18):4255–4277.
- Mahowald, N., Lo, F., Zheng, Y., Harrison, L., Funk, C., Lombardozi, D., and Goodale, C.** (2016). Projections of leaf area index in earth system models. *Earth System Dynamics*, 7(1):211–229.
- Martinuzzi, F., Mahecha, M. D., Camps-Valls, G., Montero, D., Williams, T., and Mora, K.** (2023). Learning extreme vegetation response to climate forcing: A comparison of recurrent neural network architectures.
- Martinuzzi, F., Rackauckas, C., Abdelrehim, A., Mahecha, M. D., and Mora, K.** (2022). Reservoircomputing.jl: An efficient and modular library for reservoir computing models. *Journal of Machine Learning Research*, 23(288):1–8.
- Martínez-Ferrer, L., Álvaro Moreno-Martínez, Campos-Taberner, M., García-Haro, F. J., Muñoz-Mari, J., Running, S. W., Kimball, J., Clinton, N., and Camps-Valls, G.** (2022). Quantifying uncertainty in high resolution biophysical variable retrieval with machine learning. *Remote Sensing of Environment*, 280:113199.
- Marujo, R. d. F. B., Carlos, F. M., da Costa, R. W., de Souza Arcanjo, J., Fronza, J. G., Soares, A. R., de Queiroz, G. R., and Ferreira, K. R.** (2023). A reproducible and replicable approach for harmonizing landsat-8 and sentinel-2 images. *Frontiers in Remote Sensing*, 4.
- Marujo, R. F. B., Ferreira, K. R., Queiroz, G. R., Costa, R. W., Arcanjo, J. S., and Souza, R. C. M.** (2022). Generating analysis ready data collections for brazil. In *IGARSS 2022–2022 IEEE International Geoscience and Remote Sensing Symposium*. IEEE.
- Mathieu, P.-P. and O’Neill, A.** (2008). Data assimilation: From photon counts to earth system forecasts. *Remote Sensing of Environment*, 112(4):1258–1267. Remote Sensing Data Assimilation Special Issue.

- Maxant, J., Braun, R., Caspard, M., and Clandillon, S. (2022). ExtractEO, a pipeline for disaster extent mapping in the context of emergency management. *Remote Sensing*, 14(20):5253.
- Meyer, H. and Pebesma, E. (2021). Predicting into unknown space? Estimating the area of applicability of spatial prediction models. *Methods in Ecology and Evolution*, 12(9):1620–1633.
- Meyer, H., Reudenbach, C., Hengl, T., Katurji, M., and Nauss, T. (2018). Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation. *Environmental Modelling & Software*, 101:1–9.
- Michel, J., Vinasco-Salinas, J., Inglada, J., and Hagolle, O. (2022). SEN2VENUS, a Dataset for the Training of Sentinel-2 Super-Resolution Algorithms. *Data*, 7(7).
- Mila, C., Mateu, J., Pebesma, E., and Meyer, H. (2022). Nearest neighbour distance matching Leave-One-Out Cross-Validation for map validation. *Methods in Ecology and Evolution*, 13(6):1304–1316.
- Montero, D. (2021). eemont: A python package that extends google earth engine. *Journal of Open Source Software*, 6(62):3168.
- Montero, D., Aybar, C., Ji, C., Kraemer, G., Söchtling, M., Teber, K., and Mahecha, M. D. (2024). On-demand earth system data cubes.
- Montero, D., Aybar, C., Mahecha, M. D., Martinuzzi, F., Söchtling, M., and Wieneke, S. (2023). A standardized catalogue of spectral indices to advance the use of remote sensing in Earth system research. *Scientific Data*, 10(1).
- Moreno-Martínez, Á., Izquierdo-Verdiguier, E., Maneta, M. P., Camps-Valls, G., Robinson, N., Muñoz-Mari, J., Sedano, F., Clinton, N., and Running, S. W. (2020). Multispectral high resolution sensor fusion for smoothing and gap-filling in the cloud. *Remote Sensing of Environment*, 247:111901.
- Musgrave, K., Longie, S. J., and Lim, S.-N. (2020). *Pytorch metric learning*. ArXiv, abs/2008.09164.
- Nativi, S., Mazzetti, P., and Craglia, M. (2017). A view-based model of data-cube to support big earth data systems interoperability. *Big Earth Data*, 1(1–2):75–99.
- Nikolakopoulos, K. G. (2008). Comparison of Nine Fusion Techniques for Very High Resolution Data. *Photogrammetric Engineering & Remote Sensing*, 74(5):647–659.
- Nikparvar, B. and Thill, J.-C. (2021). Machine learning of spatial data. *ISPRS International Journal of Geo-Information*, 10(9):600.
- Oscó, L. P., Wu, Q., de Lemos, E. L., Gonçalves, W. N., Ramos, A. P. M., Li, J., and Marcato, J. (2023). The segment anything model (SAM) for remote sensing applications: From zero to one shot. *International Journal of Applied Earth Observation and Geoinformation*, 124:103540.
- Oyama, N., Ishizaki, N. N., Koide, S., and Yoshida, H. (2023). Deep generative model super-resolves spatially correlated multiregional climate data. *Scientific Reports*, 13(1).
- Pabon-Moreno, D. E., Migliavacca, M., Reichstein, M., and Mahecha, M. D. (2022). On the potential of sentinel-2 for estimating gross primary production. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–12.
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library.
- Pebesma, E. and Bivand, R. (2023). *Spatial Data Science: With applications in R*. Chapman and Hall/CRC, London.
- Peng, J., Albergel, C., Balenzano, A., Brocca, L., Cartus, O., Cosh, M. H., Crow, W. T., Dabrowska-Zielinska, K., Dadson, S., Davidson, M. W., et al. (2021). A roadmap for high-resolution satellite soil moisture applications—confronting product characteristics with user requirements. *Remote Sensing of Environment*, 252:112162.
- Persello, C., Wegner, J. D., Hänsch, R., Tuia, D., Ghamisi, P., Koeva, M., and Camps-Valls, G. (2022). Deep learning and earth observation to support the sustainable development goals: Current approaches, open challenges, and future opportunities. *IEEE Geoscience and Remote Sensing Magazine*, 10(2):172–200.
- Ploton, P., Mortier, F., Réjou-Méchain, M., Barbier, N., Picard, N., Rossi, V., Dormann, C., Cornu, G., Viennois, G., Bayol, N., et al. (2020). Spatial validation reveals poor predictive performance of large-scale ecological mapping models. *Nature communications*, 11(1):1–11.
- Pérez-García, F., Sparks, R., and Ourselin, S. (2021). Torchio: A python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. *Computer Methods and Programs in Biomedicine*, 208:106236.
- Rackauckas, C., Innes, M., Ma, Y., Bettencourt, J., White, L., and Dixit, V. (2019). Diffeqflux.jl - a julia library for neural differential equations.
- Raspaud, M., Hoese, D., Lahtinen, P., Holl, G., Finkensieper, S., Proud, S., Dybbroe, A., Meraner, A., and Strandgren, J. (2023). pytrol/satpy: Version 0.44.0 (2023/10/17).
- Ravi, N., Gabeur, V., Hu, Y.-T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K. V., Carion, N., Wu, C.-Y., Girshick, R., Dollár, P., and Feichtenhofer, C. (2024). Sam 2: Segment anything in images and videos. *arXiv*.
- Razzak, M. T., Mateo-García, G., Lecuyer, G., Gómez-Chova, L., Gal, Y., and Kalaitzis, F. (2023). Multi-spectral multi-image super-resolution of Sentinel-2 with radiometric consistency losses and its effect on building delineation. *ISPRS Journal of Photogrammetry and Remote Sensing*, 195:1–13.
- Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., and Carvalhais, N. (2019). Deep learning and process understanding for data-driven Earth system science. *Nature*, 566(7743):195–204.

- Ren, H., Cromwell, E., Kravitz, B., and Chen, X. (2022). Technical note: Using long short-term memory models to fill data gaps in hydrological monitoring networks. *Hydrology and Earth System Sciences*, 26(7):1727–1743.
- Requena-Mesa, C., Benson, V., Reichstein, M., Runge, J., and Denzler, J. (2021). EarthNet2021: A large-scale dataset and challenge for Earth surface forecasting as a guided video prediction task. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.
- Richiardi, C., Blonda, P., Rana, F. M., Santoro, M., Tarantino, C., Vicario, S., and Adamo, M. (2021). A revised snow cover algorithm to improve discrimination between snow and clouds: A case study in gran paradiso national park. *Remote Sensing*, 13(10):1957.
- Rivest, S., Bédard, Y., Proulx, M.-J., Nadeau, M., Hubert, F., and Pastor, J. (2005). SOLAP technology: Merging business intelligence with geospatial technology for interactive spatio-temporal exploration and analysis of data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 60(1):17–33.
- Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Aroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., et al. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8):913–929.
- Rocklin, M. (2015). Dask: Parallel computation with blocked algorithms and task scheduling. In Huff, K. and Bergstra, J., editors, *Proceedings of the 14th Python in Science Conference*, pages 130–136.
- Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2021). High-resolution image synthesis with latent diffusion models.
- Roy, D., Zhang, H., Ju, J., Gomez-Dans, J., Lewis, P., Schaaf, C., Sun, Q., Li, J., Huang, H., and Kovalsky, V. (2016). A general method to normalize Landsat reflectance data to nadir BRDF adjusted reflectance. *Remote Sensing of Environment*, 176: 255–271.
- Runge, J., Bathiany, S., Bollt, E., Camps-Valls, G., Coumou, D., Deyle, E., Glymour, C., Kretschmer, M., Mahecha, M. D., Muñoz-Marí, J., van Nes, E. H., Peters, J., Quax, R., Reichstein, M., Scheffer, M., Schölkopf, B., Spirites, P., Sugihara, G., Sun, J., Zhang, K., and Zscheischler, J. (2019). Inferring causation from time series in Earth system sciences. *Nature Communications*, 10(1).
- Rüttgers, M., Lee, S., Jeon, S., and You, D. (2019). Prediction of a typhoon track using a generative adversarial network and satellite images. *Scientific Reports*, 9(1).
- Rußwurm, M., Klemmer, K., Rolf, E., Zbinden, R., and Tuia, D. (2023). Geographic location encoding with spherical harmonics and sinusoidal representation networks.
- Santos, L., Ferreira, K. R., Picoli, M., and Camara, G. (2019). Self-organizing maps in earth observation data cubes analysis. In *Advances in Intelligent Systems and Computing*, pages 70–79. Springer International Publishing.
- Schramm, M., Pebesma, E., Milenković, M., Foresta, L., Dries, J., Jacob, A., Wagner, W., Mohr, M., Neteler, M., Kadunc, M., et al. (2021). The openeo api—harmonising the use of earth observation cloud services using virtual data cube functionalities. *Remote Sensing*, 13(6):1125.
- Schubert, E. and Gertz, M. (2018). Numerically Stable Parallel Computation of (Co-)Variance. In *Proceedings of the 30th International Conference on Scientific and Statistical Database Management*, Ssdbm '18, pages 1–12, New York, NY, USA. Association for Computing Machinery.
- Shang, R. and Zhu, Z. (2019). Harmonizing landsat 8 and sentinel-2: A time-series-based reflectance adjustment approach. *Remote Sensing of Environment*, 235:111439.
- Sherstinsky, A. (2020). Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network. *Physica D: Nonlinear Phenomena*, 404:132306.
- Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-k., and Woo, W.-c. (2015). Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting.
- Simmons, A., Fellous, J.-L., Ramaswamy, V., Trenberth, K., Asrar, G., Balmaseda, M., Burrows, J. P., Ciais, P., Drinkwater, M., Friedlingstein, P., et al. (2016). Observation and integrated Earth-system science: A roadmap for 2016–2025. *Advances in Space Research*, 57(10):2037–2103.
- Simoes, R., Camara, G., Queiroz, G., Souza, F., Andrade, P. R., Santos, L., Carvalho, A., and Ferreira, K. (2021a). Satellite image time series analysis for big earth observation data. *Remote Sensing*, 13(13):2428.
- Simoes, R., de Souza, F. C., Zaglia, M., de Queiroz, G. R., dos Santos, R. D. C., and Ferreira, K. R. (2021b). Rstac: An R Package to Access Spatiotemporal Asset Catalog Satellite Imagery. In *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*. Ieee.
- Sippel, S., Reichstein, M., Ma, X., Mahecha, M. D., Lange, H., Flach, M., and Frank, D. (2018). Drought, heat, and the carbon cycle: a review. *Current Climate Change Reports*, 4:266–286.
- Siqueira, A., Tadono, T., Rosenqvist, A., Lacey, J., Lewis, A., Thankappan, M., Szantoi, Z., Goryl, P., Labahn, S., Ross, J., Hosford, S., and Mecklenburg, S. (2019). CEOS analysis ready data for land – an overview on the current and future work. In *IGARSS 2019–2019 IEEE International Geoscience and Remote Sensing Symposium*. IEEE.
- Skakun, S., Wevers, J., Brockmann, C., Doxani, G., Aleksandrov, M., Batič, M., Frantz, D., Gascon, F., Gómez-Chova, L., Hagolle, O., López-Puigdollers, D., Louis, J., Lubej, M., Mateo-García, G., Osman, J., Peressutti, D., Pflug, B., Puc, J., Richter, R., Roger, J.-C., Scaramuzza, P., Vermote, E., Vesel, N., Zupanc, A., and Žust, L. (2022). Cloud Mask Intercomparison eXercise (CMIX): An evaluation of cloud masking algorithms for Landsat 8 and Sentinel-2. *Remote Sensing of Environment*, 274:112990.

- Snow, A. D., Taves, M., BENR0, Cook, J., SlapDrone, Mussab Abdalla, Rambaud Pierrick, and Bell, R. (2023). *corteva/geocube: 0.4.2 release*.
- Snyder, J. P. and Voxland, P. M. (1989). *An album of map projections*. US Geological Survey.
- Söchting, M., Mahecha, M. D., Montero, D., and Scheuermann, G. (2023). *Lexcube: Interactive visualization of large earth system data cubes*. IEEE Computer Graphics and Applications, pages 1–13.
- Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.
- Steed, C. A., Evans, K. J., Harney, J. F., Jewell, B. C., Shipman, G., Smith, B. E., Thornton, P. E., and Williams, D. N. (2014). Web-based visual analytics for extreme scale climate science. In *2014 IEEE International Conference on Big Data (Big Data)*, pages 383–392. IEEE.
- Stern, C., Abernathy, R., Hamman, J., Wegener, R., Lepore, C., Harkins, S., and Merose, A. (2022). Pangeo forge: Crowdsourcing analysis-ready, cloud optimized data production. *Frontiers in Climate*, 3.
- Stewart, A. J., Robinson, C., Corley, I. A., Ortiz, A., Lavista Ferres, J. M., and Banerjee, A. (2022). TorchGeo: Deep learning with geospatial data. In *Proceedings of the 30th International Conference on Advances in Geographic Information Systems, SIGSPATIAL '22*, pages 1–12, Seattle, Washington. Association for Computing Machinery.
- Storch, H. v., Zwiers, F., and Livezey, R. (2000). Statistical analysis in climate research. *Nature*, 404(6778):544.
- Sturm, L. (2023). Analyzing earth data in deepesdl: A practical guide to cloud-based multivariate analyses on data cubes.
- Sudmanns, M., Augustin, H., Killough, B., Giuliani, G., Tiede, D., Leith, A., Yuan, F., and Lewis, A. (2022). Think global, cube local: an Earth Observation Data Cube's contribution to the Digital Earth vision. *Big Earth Data*, pages 1–29.
- Sudmanns, M., Tiede, D., Lang, S., Bergstedt, H., Trost, G., Augustin, H., Baraldi, A., and Blaschke, T. (2020). Big Earth data: disruptive changes in Earth observation data management and analysis? *International Journal of Digital Earth*, 13(7):832–850.
- Sun, Z., Sandoval, L., Crystal-Ornelas, R., Mousavi, S. M., Wang, J., Lin, C., Cristea, N., Tong, D., Carande, W. H., Ma, X., et al. (2022). A review of earth artificial intelligence. *Computers & Geosciences*, page 105034.
- Sweet, L.-b., Müller, C., Anand, M., and Zscheischler, J. (2023). Cross-validation strategy impacts the performance and interpretation of machine learning models. *Artificial Intelligence for the Earth Systems*, 2(4).
- Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic geography*, 46(sup1):234–240.
- Tuia, D., Schindler, K., Demir, B., Camps-Valls, G., Zhu, X. X., Kochupillai, M., Dzeroski, S., van Rijn, J. N., Hoos, H. H., Del Frate, F., et al. (2023). Artificial intelligence to advance Earth observation: a perspective. *arXiv preprint arXiv:2305.08413*.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need.
- Villamizar, M., Castro, H., Ariza-Porras, C., Mancipe, M. P., Cabrera, S., Pachon, I., Ramirez, S., Fonseca, D., Lozano-Rivera, P., Cabrera, E., and Becerra, M. T. (2018). Scaling the colombian data cube using a distributed architecture. In *IGARSS 2018–2018 IEEE International Geoscience and Remote Sensing Symposium*. IEEE.
- Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17: 261–272.
- Vogeler, J. C., Braaten, J. D., Slesak, R. A., and Falkowski, M. J. (2018). Extracting the full value of the Landsat archive: Inter-sensor harmonization for the mapping of Minnesota forest canopy cover (1973–2015). *Remote Sensing of Environment*, 209: 363–374.
- Wadoux, A. M.-C., Heuvelink, G. B., De Bruin, S., and Brus, D. J. (2021). Spatial cross-validation is not the right way to evaluate map accuracy. *Ecological Modelling*, 457:109692.
- Walther, S., Besnard, S., Nelson, J. A., El-Madany, T. S., Migliavacca, M., Weber, U., Carvalhais, N., Ermida, S. L., Brümmer, C., Schrader, F., Prokushkin, A. S., Panov, A. V., and Jung, M. (2022). Technical note: A view from space on global flux towers by MODIS and Landsat: the FluxnetEO data set. *Biogeosciences*, 19(11):2805–2840.
- Wang, Y., Köhler, P., Braghiere, R. K., Longo, M., Doughty, R., Bloom, A. A., and Frankenberg, C. (2022). GriddingMachine, a database and software for earth system modeling at global and regional scales. *Scientific Data*, 9(1).
- Welford, B. P. (1962). Note on a Method for Calculating Corrected Sums of Squares and Products. *Technometrics*, 4(3):419–420.
- Wickham, H. (2011). The Split-Apply-Combine Strategy for Data Analysis. *Journal of Statistical Software*, 40(1):1–29.
- Wieczorek, M. A. and Meschede, M. (2018). SHTools: Tools for Working with Spherical Harmonics. *Geochemistry, Geophysics, Geosystems*, 19(8):2574–2592.
- Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J. -W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J., Groth, P., Goble, C., Grethe, J. S., Heringa, J., 't Hoen, P. A., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., and Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1).

- Wu, Q. and Osco, L. P.** (2023). samgeo: A python package for segmenting geospatial data with the segment anything model (sam). *Journal of Open Source Software*, 8(89):5663.
- Wulder, M. A., Hilker, T., White, J. C., Coops, N. C., Masek, J. G., Pflugmacher, D., and Crevier, Y.** (2015). Virtual constellations for global terrestrial monitoring. *Remote Sensing of Environment*, 170:62–76.
- Yates, K. L., Bouchet, P. J., Caley, M. J., Mengersen, K., Randin, C. F., Parnell, S., Fielding, A. H., Bamford, A. J., Ban, S., Barbosa, A. M., et al.** (2018). Outstanding challenges in the transferability of ecological models. *Trends in ecology & evolution*, 33(10):790–802.
- Yu, X., Orth, R., Reichstein, M., Bahn, M., Klosterhalfen, A., Knohl, A., Koepsch, F., Migliavacca, M., Mund, M., Nelson, J. A., Stocker, B. D., Walther, S., and Bastos, A.** (2022). Contrasting drought legacy effects on gross primary productivity in a mixed versus pure beech forest. *Biogeosciences*, 19(17):4315–4329.
- Zellner, P. J., Claus, M., Dolezalova, T., Balogun, R. O., Eberle, J., Hodam, H., Eckardt, R., Meibl, S., Jacob, A., and Anghilea, A.** (2024). Mooc cubes and clouds - cloud native open data sciences for earth observation. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLVIII-4/W12-2024:157–162.
- Zscheischler, J., Martius, O., Westra, S., Bevacqua, E., Raymond, C., Horton, R. M., van den Hurk, B., AghaKouchak, A., Jézéquel, A., Mahecha, M. D., et al.** (2020). A typology of compound weather and climate events. *Nature reviews earth & environment*, 1(7):333–347.
- Zuefle, A., Wessels, K., and Pfoser, D.** (2021). Mining high resolution earth observation data cubes. In *17th International Symposium on Spatial and Temporal Databases*. ACM.

Cite this article: Montero D, Kraemer G, Anghilea A, Aybar C, Brandt G, Camps-Valls G, Cremer F, Flik I, Gans F, Habershon S, Ji C, Kattenborn T, Martínez-Ferrer L, Martinuzzi F, Reinhardt M, Söchting M, Teber K and Mahecha MD (2024). Earth System Data Cubes: Avenues for advancing Earth system research. *Environmental Data Science*, 3: e27. doi:10.1017/eds.2024.22