Some thoughts on ethical implications of digital twin technology in Earth Sciences

Ramon Carbonell, CSIC-GEO3BCN

Board on ethics:

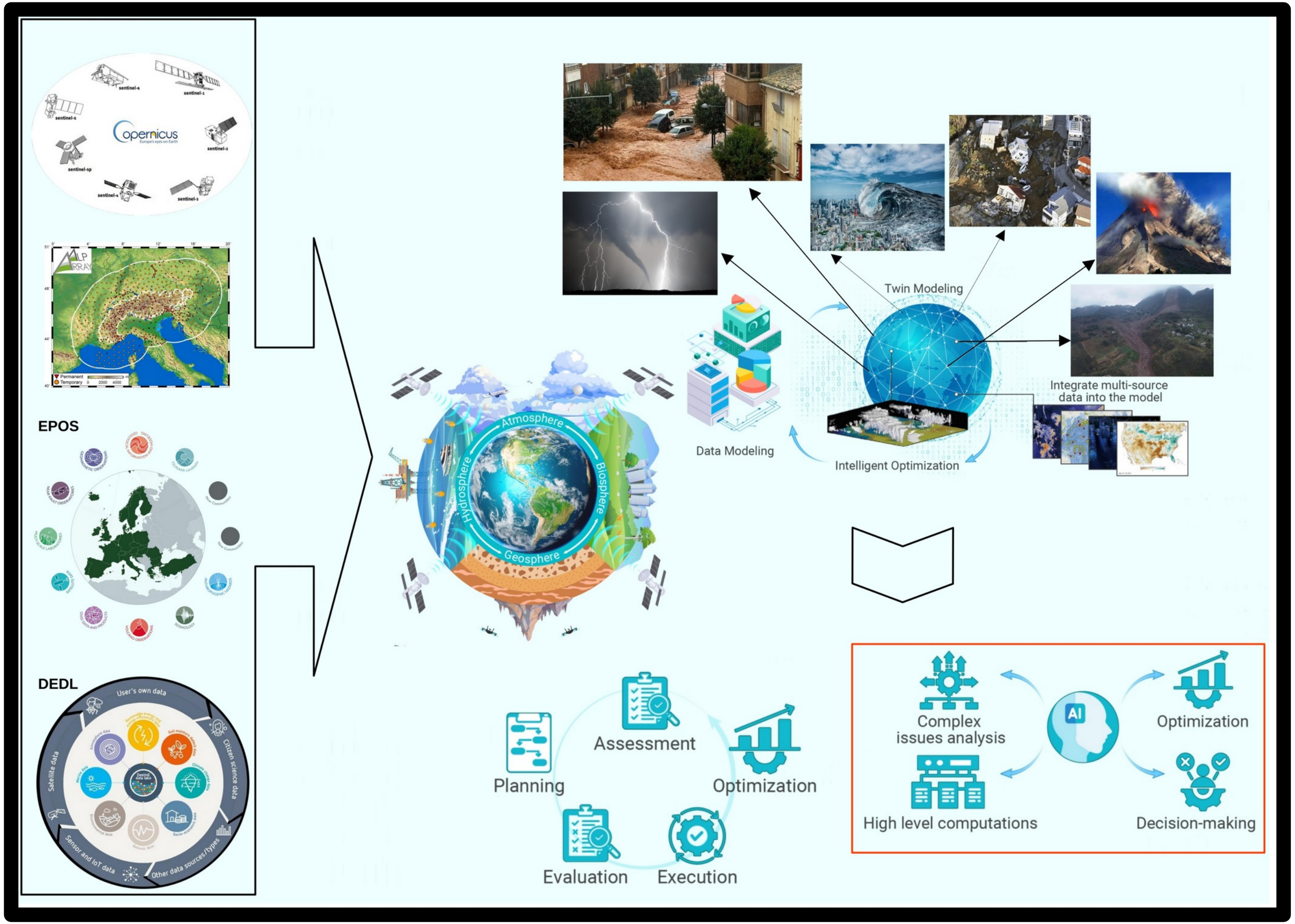
Marisol Monterrubio, BSC
Cedric Bhihe, BSC
Steven Gibbons, NGI
Flavio Cannavo, INGV
Jörn Behrens, Univ. Hamburg

11th. Geo-INQUIRE online Seminar "Curiosity-driven and cross disciplinary applications of AI"

Thursday 22sd. May 2025.

This project has received funding from the European Union's Horizon research and innovation programme under the grant agreement No 101058129

Concept: Digital Twin Technology



DT-GEO

Motivation:

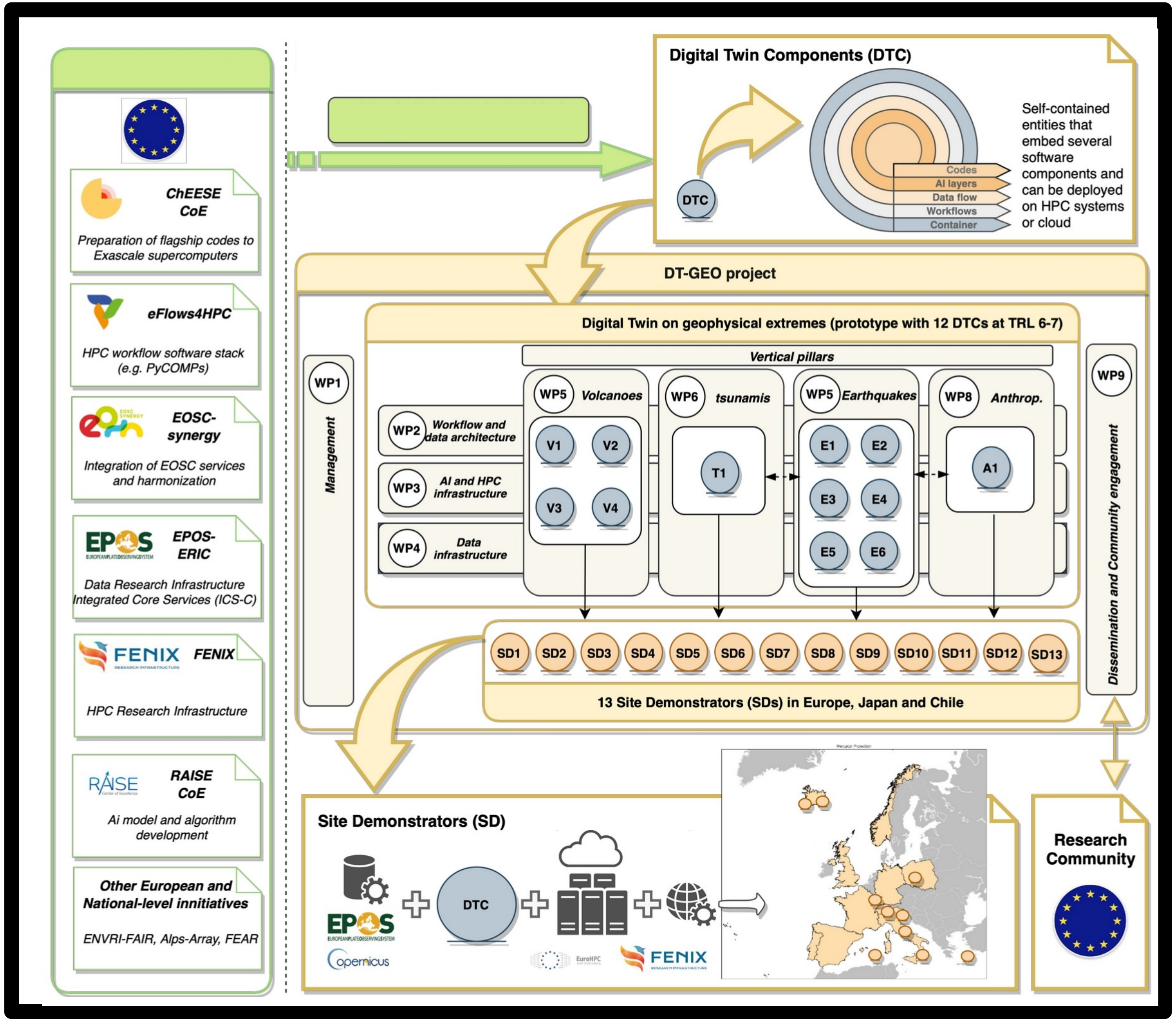
“DT-GEO’s prototypes ***aim to be used to increase human well-being by facilitating risk assessment, and forecasting possible behavior in the case of extreme geophysical events***. Produce data informed responses to what if questions.

These tools involve to some extent AI/ML approaches. Therefore the motivation is:
To what degree the AI/ML schemes used influence the risks assessments and forecasts? And, what are the implications when these schemes *make biased or unfair decisions?*

The final aim is to build trust in this technology so that it is used in a safe way (do good) compliant with the law, including a maximum respect of fundamental rights.

Concept:

DT-GEO



ChEESE CoE

Preparation of flagship codes to Exascale supercomputers

eFlows4HPC

HPC workflow software stack (e.g. PyCOMPs)

EOSC-synergy

Integration of EOSC services and harmonization

EPOS-ERIC

Data Research Infrastructure Integrated Core Services (ICS-C)

FENIX

HPC Research Infrastructure

RAISE CoE

Ai model and algorithm development

Other European and National-level innitiatives

ENVRI-FAIR, Alps-Array, FEAR

Digital Twin Components (DTC)

Self-contained entities that embed several software components and can be deployed on HPC systems or cloud

DT-GEO project

Digital Twin on geophysical extremes (prototype with 12 DTCs at TRL 6-7)

13 Site Demonstrators (SDs) in Europe, Japan and Chile

Site Demonstrators (SD)

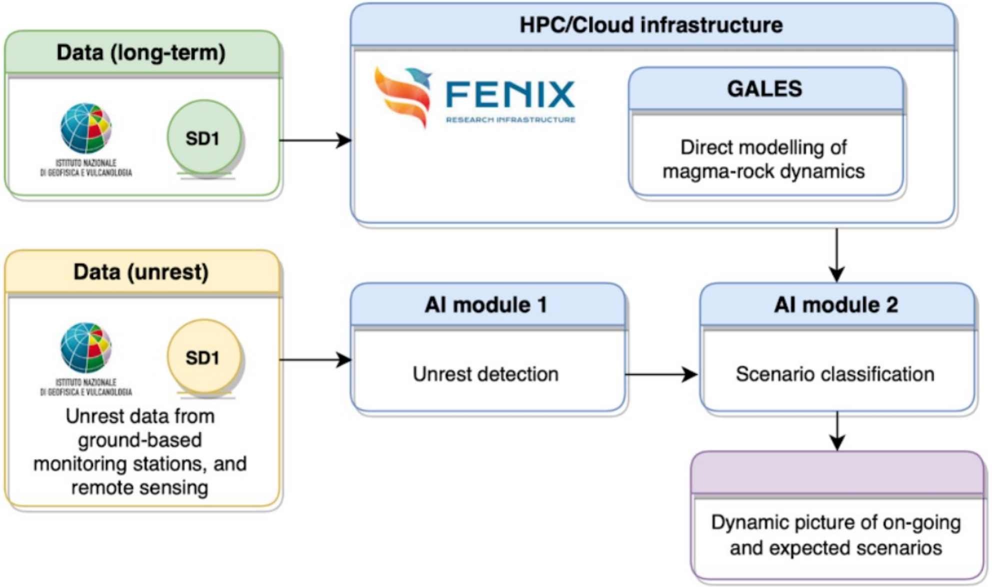
Research Community

12 Digital Twin Components (DTCs): Use of AI/ML modules

DTC	Code	Hazard	Name	Target TRL	Site Demonstrator
1	**DTC-V1**	Volcano	**Volcanic unrest dynamics**	6	SD1
2	**DTC-V2**		**Volcanic ash clouds and deposition**	7	SD2
3	**DTC-V3**		**Lava flows**	6	SD1, SD3
4	DTC-V4		Volcanic gas dispersal and deposition	7	SD3
5	**DTC-T1**	Tsunami	**Probabilistic Tsunami Forecasting (PTF)**	7	SD4, SD5, SD6, SD7
6	DTC-E1	Earthquake	Probabilistic Seismic Hazard and Risk Assessment	7	SD8
7	**DTC-E2**		**Earthquake short-term forecasting**	7	SD8, SD9
8	**DTC-E3**		**Tomography and Ground Motion Models (GMM)**	7	SD8, SD9
9	DTC-E4		Fault rupture forecasting	7	SD9, SD10
10	**DTC-E5**		**Tomography and shaking simulation**	6	SD8, SD11
11	**DTC-E6**		**Rapid event and shaking characterization**	7	SD8
12	**DTC-A1**	Anthropogenic	**Anthropogenic geophysical extreme forecasting**	6	SD12, SD13

DTC-V1:
AI/ML use & location

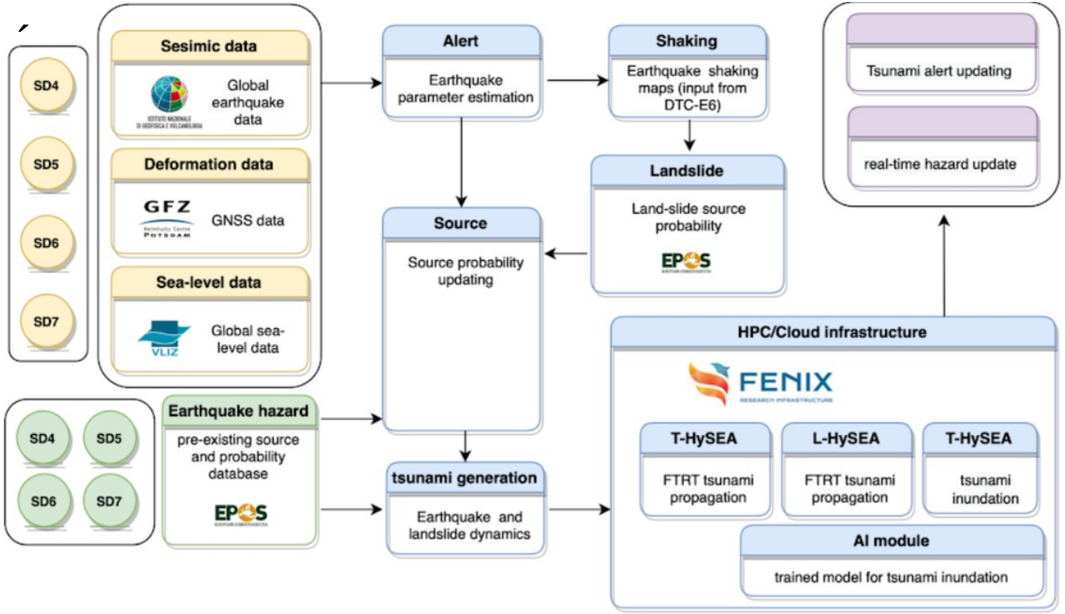
Volcanic Unrest dynamics


Data (long-term)
ISTITUTO NAZIONALE DI GEOFISICA E VULCANOLOGIA
SD1
HPC/Cloud infrastructure
FENIX
RESEARCH INFRASTRUCTURE
GALES
Direct modelling of magma-rock dynamics
Data (unrest)
ISTITUTO NAZIONALE DI GEOFISICA E VULCANOLOGIA
SD1
Unrest data from ground-based monitoring stations, and remote sensing
AI module 1
Unrest detection
AI module 2
Scenario classification
Dynamic picture of on-going and expected scenarios

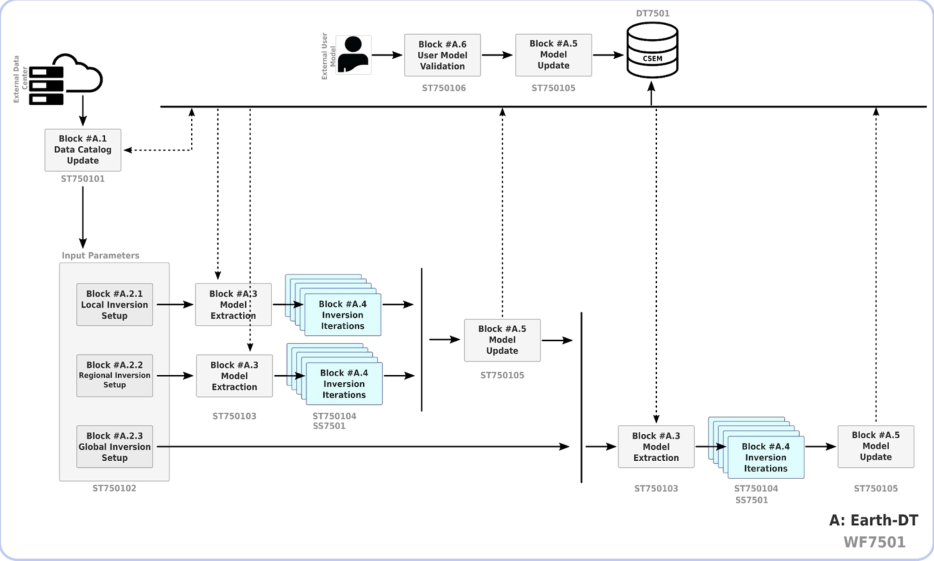
AI/ML Location within the workflow.	ML Module's mission & funtions
AI m-1: Unrest detection	Detects anomalies on multiparameter data in real time indicative of system unrest [a rule-based classifier]. Works under human supervision
AI m-2: Scenario classification	Spatial identification of possible source of unrest-related deformation. The module uses a deep model.

DTC-T1:
AI/ML use & location

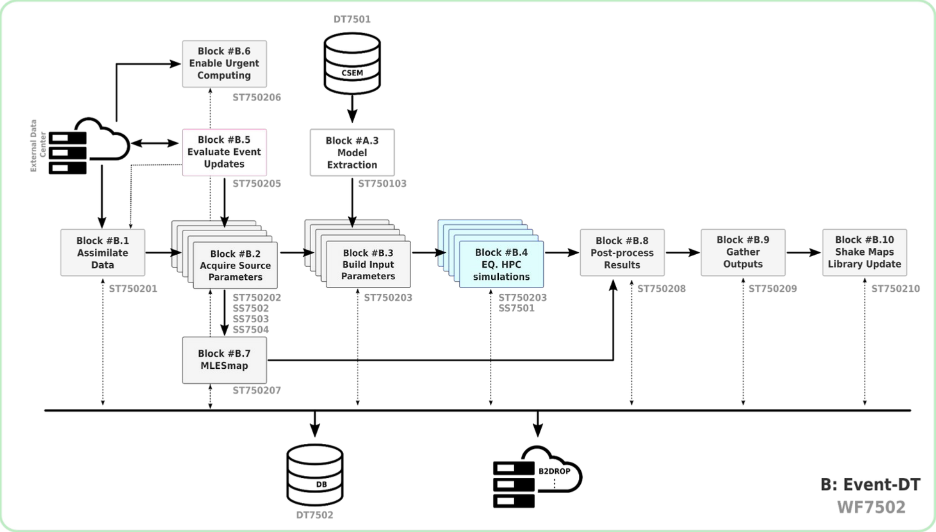
Probabilistic Tsunami Forecasting



AI/ML Location within the workflow.	ML Module's mission & funtions
AI module, trained module for Tsunami inundation emulation	Emulator based on AI models trained from previous HPC simulation. Its input is time series from the offshore simulations used in Probabilistic Tsunami Forecasting tools. *The workflow functions without the AI module but that it can be implemented to save simulation time.* Provides inundation maps for a set of test areas where the AI training has been conducted. It increase model accuracy and resolution it reduces the computational load (green AI).

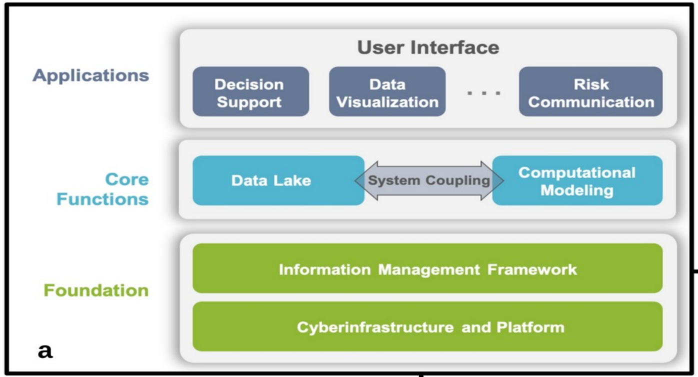

DT7501
External Data Center
External User Model
Block #A.6 User Model Validation
ST750106
Block #A.5 Model Update
ST750105
CSEM
Block #A.1 Data Catalog Update
ST750101
Input Parameters
Block #A.2.1 Local Inversion Setup
Block #A.2.2 Regional Inversion Setup
Block #A.2.3 Global Inversion Setup
ST750102
Block #A.3 Model Extraction
Block #A.3 Model Extraction
ST750103
Block #A.4 Inversion Iterations
Block #A.4 Inversion Iterations
ST750104
SS7501
Block #A.5 Model Update
ST750105
Block #A.3 Model Extraction
ST750103
Block #A.4 Inversion Iterations
ST750104
SS7501
Block #A.5 Model Update
ST750105
A: Earth-DT
WF7501

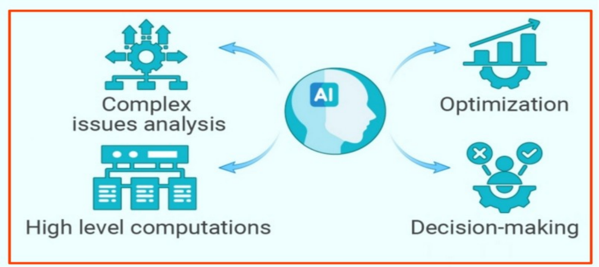
Tomography & Shaking Simulations


DT7501
CSEM
Block #B.6 Enable Urgent Computing
ST750206
External Data Center
Block #B.5 Evaluate Event Updates
ST750205
Block #A.3 Model Extraction
ST750103
Block #B.1 Assimilate Data
ST750201
Block #B.2 Acquire Source Parameters
ST750202
SS7502
SS7503
SS7504
Block #B.3 Build Input Parameters
ST750203
Block #B.4 EQ. HPC simulations
ST750203
SS7501
Block #B.8 Post-process Results
ST750208
Block #B.9 Gather Outputs
ST750209
Block #B.10 Shake Maps Library Update
ST750210
Block #B.7 MLESmap
ST750207
DB
DT7502
B2DROP
B: Event-DT
WF7502

DTC-E5: AI/ML use & location

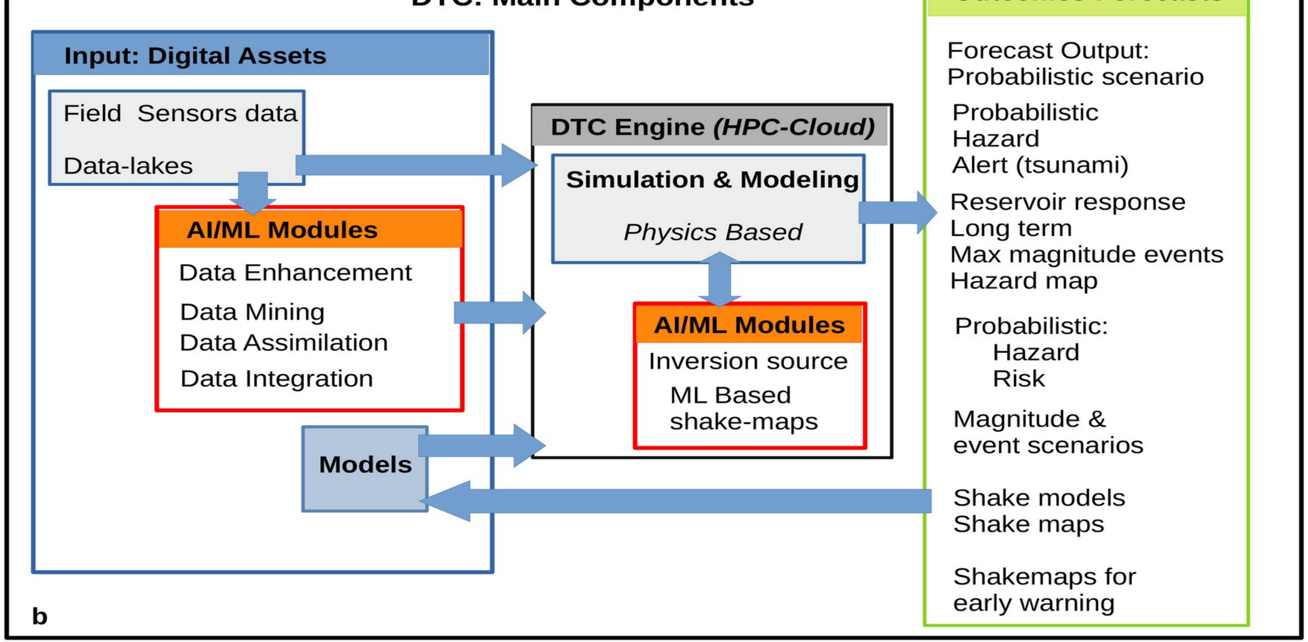
AI/ML Location within the workflow.	ML Module’s mission & funtions
MLES Map	Computes ground shaking maps based on ML models trained offline. Those inference results inform the uncertainty quantification method and also provide real-time event impact information ahead of the simulated results.
Source Parameters	Rapidly estimates the source parameters that are used for parametrising HPC simulations that provide synthetic ground motion proxies.





DTC's AI/ML Modules

Location & mission



The objectives & expectations of AI/ML in Earth System DTs [Le Mogine, 2023]:

- *Improving data content and information extraction*
- *Improving data fusion and data assimilation*
- *Improving models spatial and temporal accuracy*
- *Enabling model interconnection*
- *Accurate and trusted surrogate modeling*
- *Providing full explainability*
- *Integrating or fully relating to physics models*
- *Speeding up What-If simulations*
- *Enabling causal analysis and impact assessment*
- *Enabling straightforward, dynamic, and interactive user interfaces.*

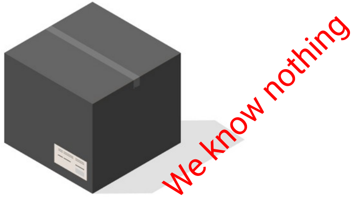
DTC’s only use ML

Risk assessment and forecasting tools (involving AI/ML) are aimed to be used to increase human well-being and do “good”. We, society, people, need to trust this technology and use it in a safe way compliant with the law, including a maximum respect of fundamental rights.

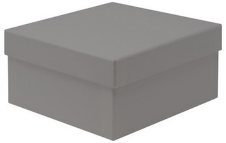
Build Trust

"Jidoka" humans give wisdom to machines

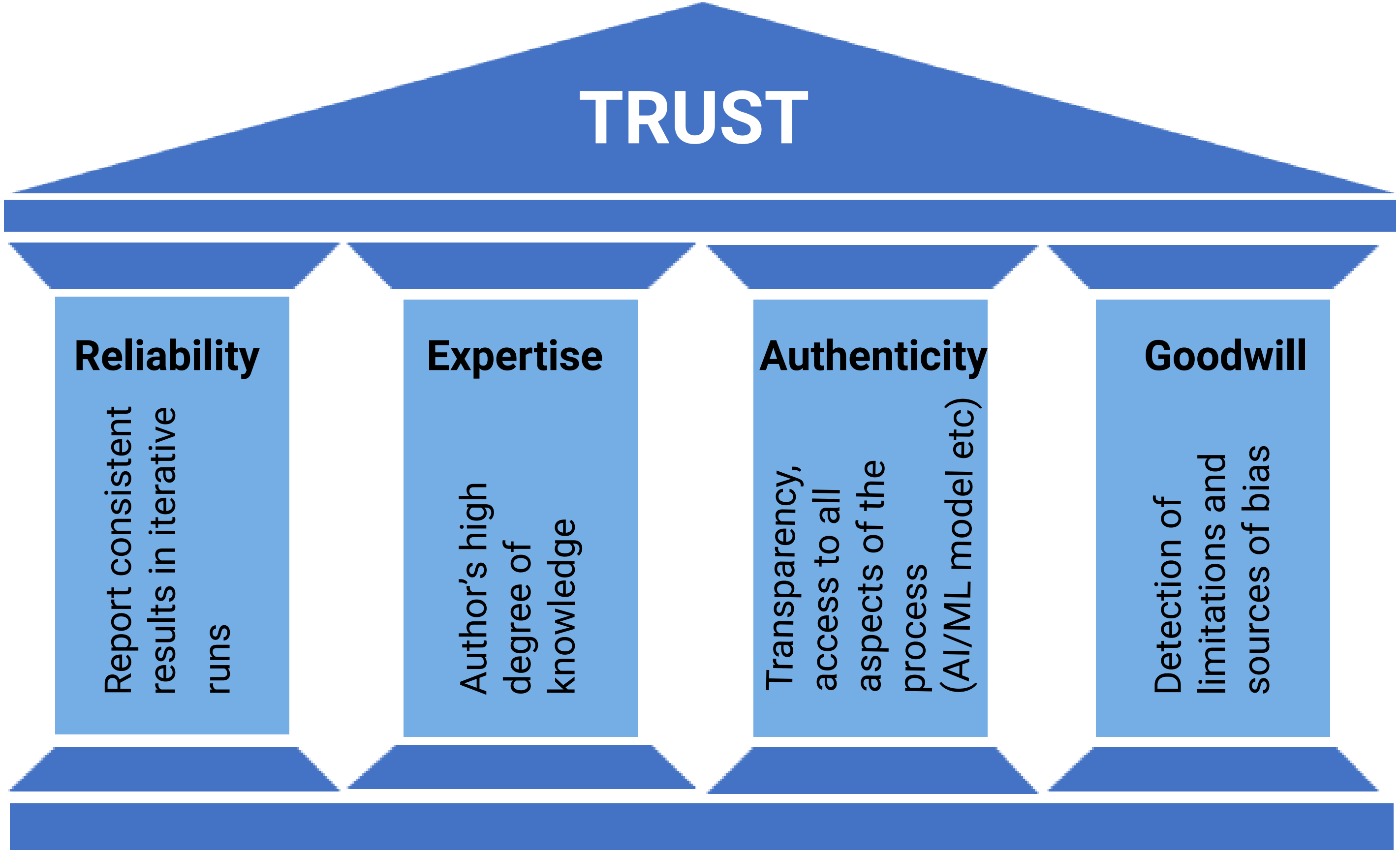
Full understanding of all the steps involved in the solution process



Steps are documented, computational process is not



Assume risks rely on the developer. **Transparency** mandatory



TRUST

Reliability

Report consistent results in iterative runs

Expertise

Author's high degree of knowledge

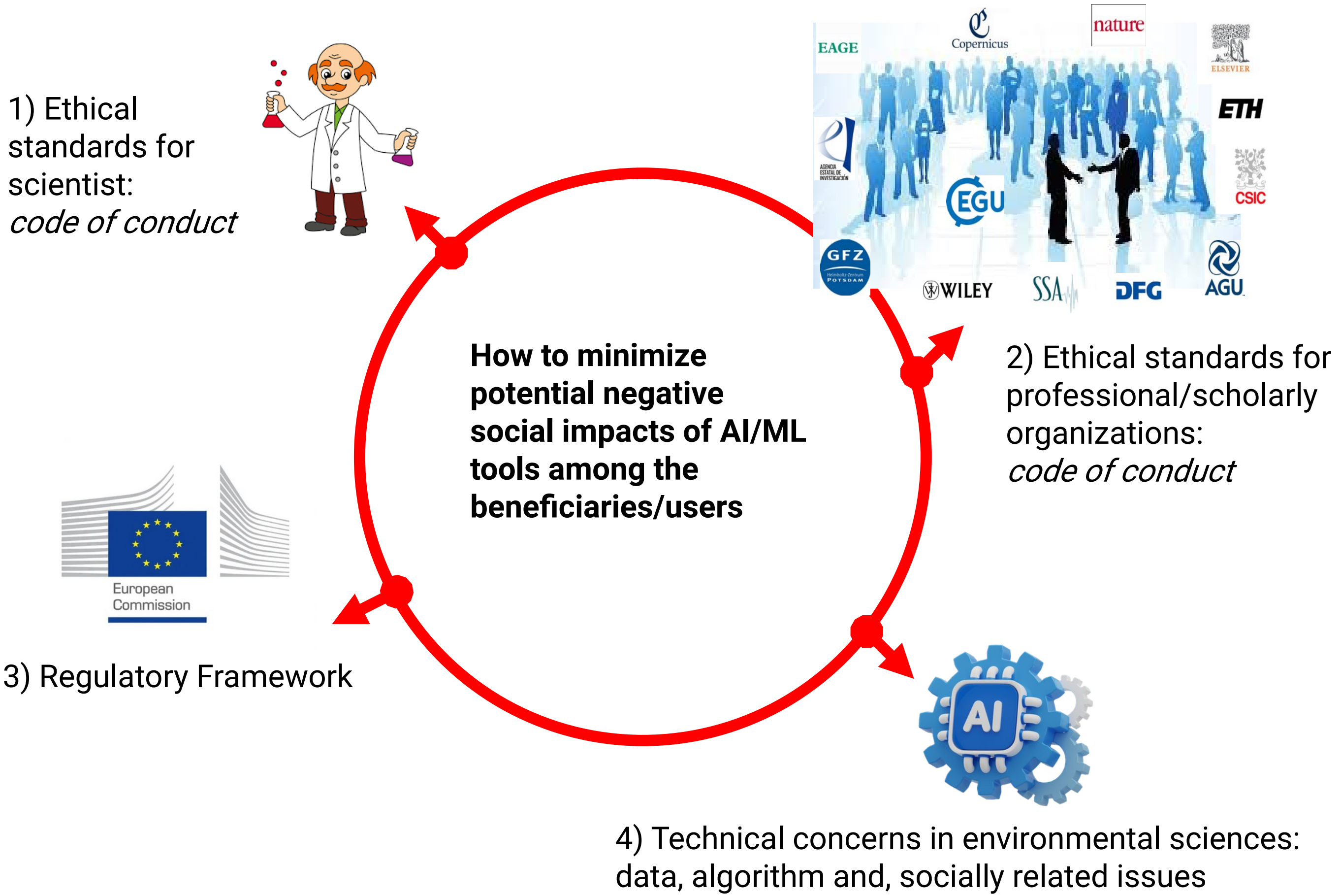
Authenticity

Transparency, access to all aspects of the process (AI/ML model etc)

Goodwill

Detection of limitations and sources of bias

Dimensions of trust building



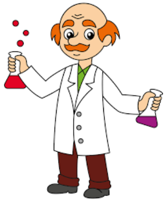
1) Ethical standards for scientist: *code of conduct*

2) Ethical standards for professional/scholarly organizations: *code of conduct*

3) Regulatory Framework

4) Technical concerns in environmental sciences: data, algorithm and, socially related issues

How to minimize potential negative social impacts of AI/ML tools among the beneficiaries/users



Researcher’s Code of conduct

1	**Transparency, Documenting, Reporting:**
	Full documentation, & reporting: identification of developers, procedures, uncertainty assessment, data and model. biases. Follow open science leading practices (FAIR, TRUST, etc). Documenting: data collection (sources, data provenance and availability, pre-proccessing); model construction, training (parameter values, etc.), validation, and results. Alerts on known biases in data. Inform on possible influences on downstream applications.
2	**Intentionality, Interpretability, Explainability, Reproducibility, and Replicability:**
	Justify the method chosen, with alternatives considered. Model specification and documentation (through its development) & evidences that the model used works as intended, Assure the replicability of the outcomes (provision of data examples).
3	**Risk, Bias, and Effects:**
	Determine and understand risks & biases & its possible mitigation. Differentiate between intended & unintended consequences & associated harms, to help manage and respond to undesired outcomes. All involved in the process (including funding agencies) share responsibility for managing risk, bias and harm and demands shared accountability.
4	**Participatory Methods:**
	Ensure: 1) consent of all individuals involved (researchers, users and communities impacted by the application of the tools). 2) voluntary & continuing consent from communities who may be impacted. 3) representation in decision-making. Research teams: designed with inclusion and diversity in mind at all stages.

Scholarly Organizations: Code of conduct

1	**Outreach, Training, and Leading Practices:**
	Enable researchers, practitioners, funding bodies, and specialized community to have awareness, understanding, and access to training. Ensuring: knowledge & hands on training. Integration of the broader community (including human-social sciences and ethics, & general public) to ensure diverse, inclusive and comprehensive contributions favoring results in high quality science and positive public impact. Provide training & access to resources and develop curricula for future specialists & decision makers. Asses tools' limitations & learn on the adverse application, miss-use of the tools causing harm.
2	**Considerations for Organizations and Institutions, Publishers, Societies, and Funders:**
	Promote and/or foster a culture around ethical practices (eg. codes of conduct) articulating values, designing governance and enforcing responsibilities.

European Commission

Regulatory Framework

European Union requirements for a trustworthy AI :

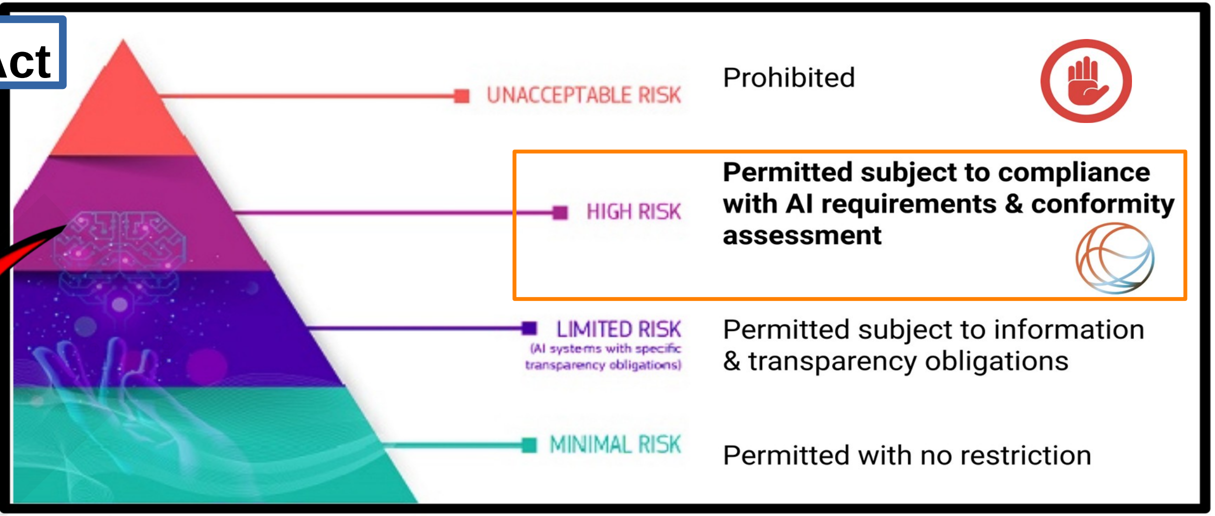
7 principals driving to increase human well-being and do “good”.

1. ***Respect for human agency*** *(“Machines can never be in full control.”)*
2. ***Technical robustness and safety*** *(“Safety risk assessment must be built in.”)*
3. ***Privacy and data governance*** *(“Citizens should keep full control over their data, including the right to see their data expunged from any AI system to the extent permitted by law.”)*
4. ***Transparency*** *(“Traceability of data and processes - viz FAIR. AI systems must be forward in identifying themselves as such. Explainability is instrumental to ensure transparency.)*
5. ***Diversity, nondiscrimination and fairness*** *(“Avoid unfair biases, ensure equal rights and opportunities.”)*
6. ***Individual, social and environmental well-being*** *(“The effects of AI systems on the natural, political and human environments should be assessed.”)*
7. ***Accountability and oversight*** *(Impact assessment must be considered at AI design phase. Evaluation and reporting of results should be continuous and redress mechanism implemented and made available to individuals.)*

EU AI Act

European Commission

Regulatory Framework



UNACCEPTABLE RISK
Prohibited
HIGH RISK
Permitted subject to compliance with AI requirements & conformity assessment
LIMITED RISK
(AI systems with specific transparency obligations)
Permitted subject to information & transparency obligations
MINIMAL RISK
Permitted with no restriction

High-risk AI/ML Systems: Actors & Obligations

Providers	Quality Management System
	Technical documentation
	Conformity assessment
	Archive automatically generated logs
	Cooperation with competent authorities
	Appointment of legal representative
Users	Monitor the operation of the AI system
	Keep the logs automatically generated

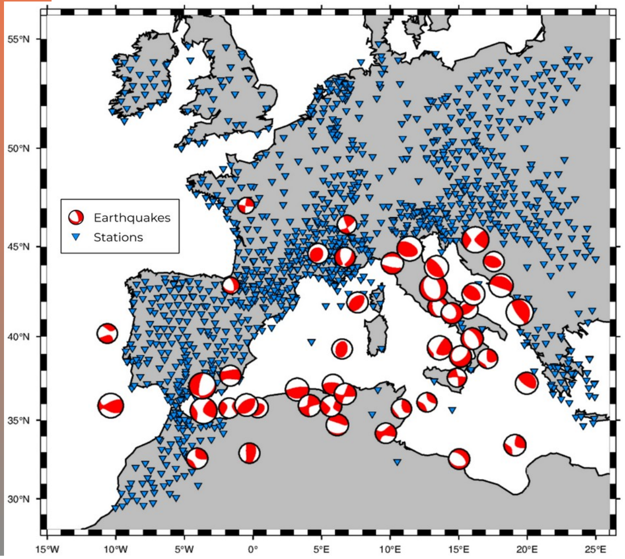
*	Appointment of national notifying authorities
*	Definition of common specification, when standards are missing
*	Conformity assessment procedures
*	EU Declaration of conformity
*	CE Marking
*	Report malfunctioning (and serious incidents)
*	EU data base registration

Technical concerns in environmental sciences

Input or data related factors	
D1	**Data representativeness:** Data is affected by spatial-temporal-spectral acquisition irregularities. Uneven sampling in any axis (space, time, frequency) might results in the data not being representative of the process addressed. Uncertainties due too sampling strategies (purposive, random sampling), uneven geographical acquisition geometries, spatial co-registration, data imbalance.
D2	**Existence of adversaries in the data:** Existence of faulty values in the input data due to natural & or harware malfunction (eg. thunderstorms → spikes).
D3	**Miss-labeling:** Iregular sensor distribution areas can have different degrees of representation (eg. overpopulated areas, or areas with complex logistics Country's science funding programs in Earth sciences, (sensor instrumentation costs play a critical role in the number and distribution of sensor hardware eg. in Seismology, Air quality measurements, etc).

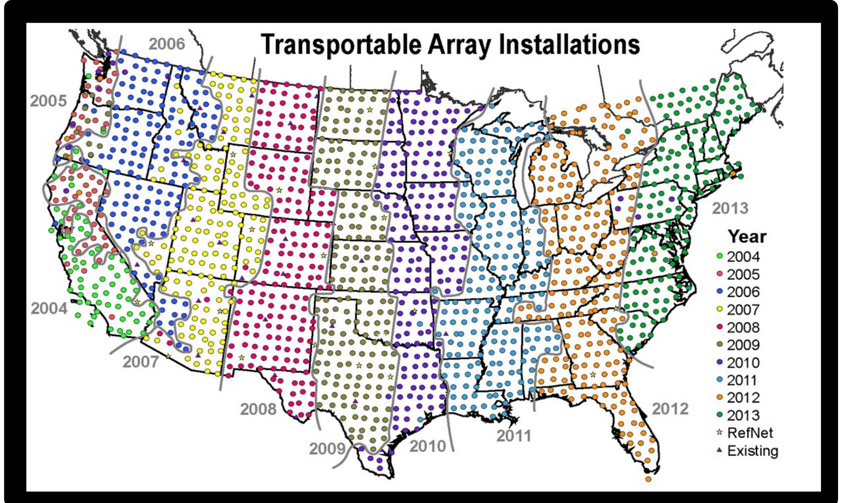
Attenuating measures: **Data QC & FAIR validation rely on data-lake providers (eg. EPOS), multiple site demonstrators**

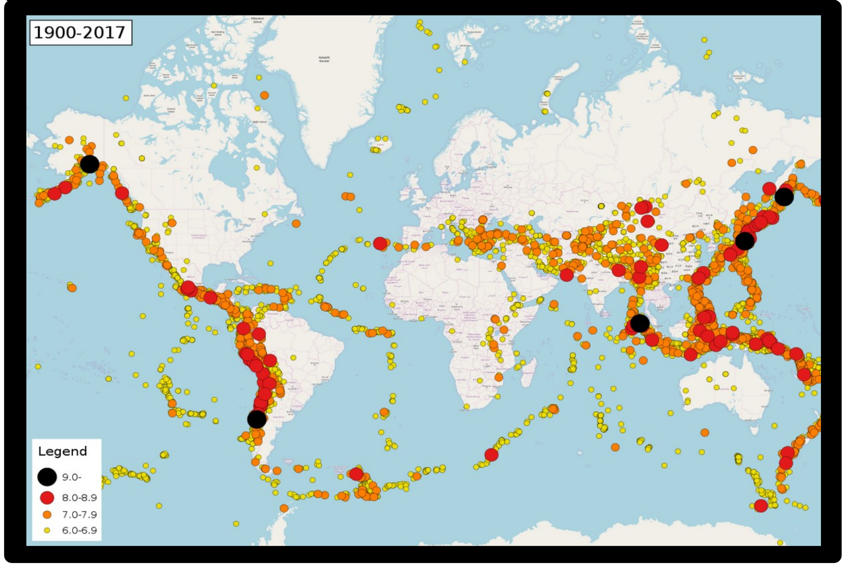
Technical concerns in environmental sciences



DTC’s rely mostly on:

- Sensor data which can be affected by heterogeneous distribution of sensors receivers and/or sources events, irregular data sampling
- Physics based



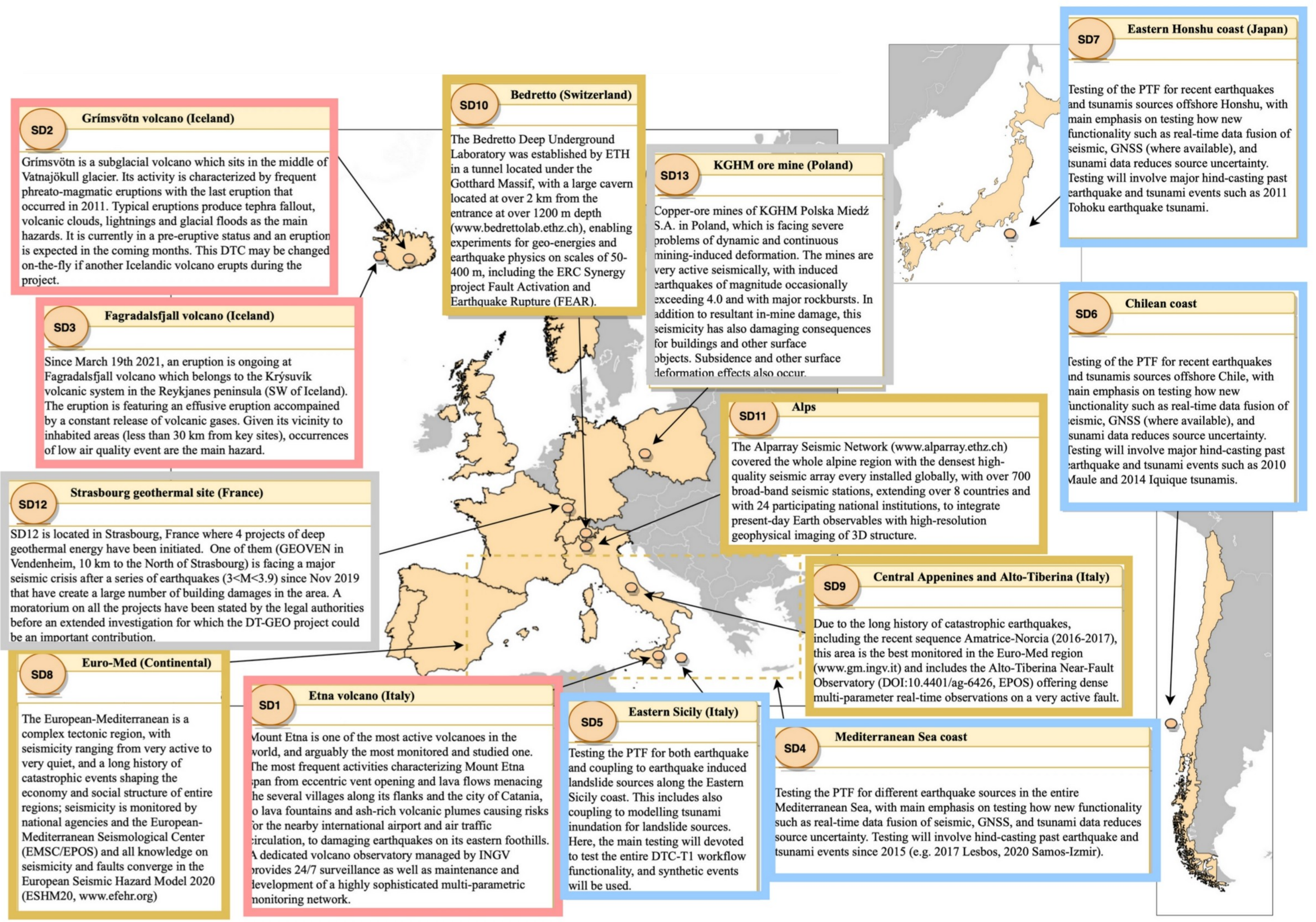


Technical concerns in environmental sciences

Processing algorithm/model related factors	
M1	**Weak and non-trustworthy model:** Models designed, trained & validated (including system's accuracy & uncertainty determined) for specific areas. Lack of model robustness can result in meaningless outcomes when the input data is not within the characteristics of the training dataset (include physics of the process and FAIR principles).
M2	**Inappropriate model design:** Designing & training the models at the specific geo-location always.
M3	**Hallucination (faking possible outcomes):** In sparse data regions models can insert faulty features if training is carried out without physically constrained data.
M4	**Algorithm learns unreliable strategies:**Bogus noisy data can contain hidden patterns which AI/ML schemes will detect generating undesired outcomes. Enable the means to potentially identify unreliable strategies before the deployment.
M5	**Training options:** Operator's decisions on data & models parameters, data provenance, resolution, preprocessing, hyper parameters choices (e.g., in random fores numbre of tress, maximal depth, etc) impact the results and drive to significantly different outcomes and implications.

Attenuating measures: **Inclusion physics based forecasts, multiple site validation, documentation (transparency), FAIR validation on digital assets,**

Technical concerns in environmental sciences
Multiple site validation



SD2 Grímsvötn volcano (Iceland)

Grímsvötn is a subglacial volcano which sits in the middle of Vatnajökull glacier. Its activity is characterized by frequent phreato-magmatic eruptions with the last eruption that occurred in 2011. Typical eruptions produce tephra fallout, volcanic clouds, lightnings and glacial floods as the main hazards. It is currently in a pre-eruptive status and an eruption is expected in the coming months. This DTC may be changed on-the-fly if another Icelandic volcano erupts during the project.

SD3 Fagradalsfjall volcano (Iceland)

Since March 19th 2021, an eruption is ongoing at Fagradalsfjall volcano which belongs to the Krýsuvík volcanic system in the Reykjanes peninsula (SW of Iceland). The eruption is featuring an effusive eruption accompained by a constant release of volcanic gases. Given its vicinity to inhabited areas (less than 30 km from key sites), occurrences of low air quality event are the main hazard.

SD12 Strasbourg geothermal site (France)

SD12 is located in Strasbourg, France where 4 projects of deep geothermal energy have been initiated. One of them (GEOVEN in Vendenheim, 10 km to the North of Strasbourg) is facing a major seismic crisis after a series of earthquakes (3<M<3.9) since Nov 2019 that have create a large number of building damages in the area. A moratorium on all the projects have been stated by the legal authorities before an extended investigation for which the DT-GEO project could be an important contribution.

SD8 Euro-Med (Continental)

The European-Mediterranean is a complex tectonic region, with seismicity ranging from very active to very quiet, and a long history of catastrophic events shaping the economy and social structure of entire regions; seismicity is monitored by national agencies and the European-Mediterranean Seismological Center (EMSC/EPOS) and all knowledge on seismicity and faults converge in the European Seismic Hazard Model 2020 (ESHM20, www.efehr.org)

SD1 Etna volcano (Italy)

Mount Etna is one of the most active volcanoes in the vorld, and arguably the most monitored and studied one. The most frequent activities characterizing Mount Etna span from eccentric vent opening and lava flows menacing he several villages along its flanks and the city of Catania, o lava fountains and ash-rich volcanic plumes causing risks for the nearby international airport and air traffic irculation, to damaging earthquakes on its eastern foothills. A dedicated volcano observatory managed by INGV rovides 24/7 surveillance as well as maintenance and levelopment of a highly sophisticated multi-parametric nonitoring network.

SD10 Bedretto (Switzerland)

The Bedretto Deep Underground Laboratory was established by ETH in a tunnel located under the Gotthard Massif, with a large cavern located at over 2 km from the entrance at over 1200 m depth (www.bedrettolab.ethz.ch), enabling experiments for geo-energies and earthquake physics on scales of 50-400 m, including the ERC Synergy project Fault Activation and Earthquake Rupture (FEAR).

SD13 KGHM ore mine (Poland)

Copper-ore mines of KGHM Polska Miedź S.A. in Poland, which is facing severe problems of dynamic and continuous mining-induced deformation. The mines are very active seismically, with induced earthquakes of magnitude occasionally exceeding 4.0 and with major rockbursts. In addition to resultant in-mine damage, this seismicity has also damaging consequences for buildings and other surface objects. Subsidence and other surface deformation effects also occur.

SD11 Alps

The Alparray Seismic Network (www.alparray.ethz.ch) covered the whole alpine region with the densest high-quality seismic array every installed globally, with over 700 broad-band seismic stations, extending over 8 countries and with 24 participating national institutions, to integrate present-day Earth observables with high-resolution geophysical imaging of 3D structure.

SD9 Central Appenines and Alto-Tiberina (Italy)

Due to the long history of catastrophic earthquakes, including the recent sequence Amatrice-Norcia (2016-2017), this area is the best monitored in the Euro-Med region (www.gm.ingv.it) and includes the Alto-Tiberina Near-Fault Observatory (DOI:10.4401/ag-6426, EPOS) offering dense multi-parameter real-time observations on a very active fault.

SD5 Eastern Sicily (Italy)

Testing the PTF for both earthquake and coupling to earthquake induced landslide sources along the Eastern Sicily coast. This includes also coupling to modelling tsunami inundation for landslide sources. Here, the main testing will devoted to test the entire DTC-T1 workflow functionality, and synthetic events will be used.

SD4 Mediterranean Sea coast

Testing the PTF for different earthquake sources in the entire Mediterranean Sea, with main emphasis on testing how new functionality such as real-time data fusion of seismic, GNSS, and tsunami data reduces source uncertainty. Testing will involve hind-casting past earthquake and tsunami events since 2015 (e.g. 2017 Lesbos, 2020 Samos-Izmir).

SD7 Eastern Honshu coast (Japan)

Testing of the PTF for recent earthquakes and tsunamis sources offshore Honshu, with main emphasis on testing how new functionality such as real-time data fusion of seismic, GNSS (where available), and tsunami data reduces source uncertainty. Testing will involve major hind-casting past earthquake and tsunami events such as 2011 Tohoku earthquake tsunami.

SD6 Chilean coast

Testing of the PTF for recent earthquakes nd tsunamis sources offshore Chile, with nain emphasis on testing how new unctionality such as real-time data fusion of eismic, GNSS (where available), and sunami data reduces source uncertainty. Testing will involve major hind-casting past arthquake and tsunami events such as 2010 Maule and 2014 Iquique tsunamis.

Technical concerns in environmental sciences

Social and Human related factors	
S1	**Implications on the C0$_2$ footprint:** High HPC demands of running DT and AI apps impacts on the carbon footprint. Research efforts are under development
S2	**Obstruct efforts of third world country's developments:** Local communities need to be involved and their necessities, concerns taken into account by the suppliers so that unintended consequences of the algorithms can be monitored and attenuated.
S3	**Lack of approval for data acquisition and/or model training:** All people involved, affected and local experts need to be participating since the design steps, as their knowledge, contacts, etc, constitute a strong link with the target scenario and a secure connection to access locally acquired data. Pursue strong public engagement in the full AI deployment process.
S4	**Vulnerability of scientific workforce:** The deployments of these technologies requires leading edge knowledge and access to super-computing infrastructures. Researchers with access to this facilities feature a clear advantage beyond the rest of the community in the field. Efforts need to be pursued to leverage access to HPC a train all research teams specially in developing countries.

Within EU & at prototype level, S2, S3, S4 factors are probably of limited significance.

DTC	Code	Hazard	Digital Twin Components *0%* *50%* *100%*	Tech	Site Demonstrator
1	**DTC-V1**	Volcano	**Volcanic unrest dynamics** 🟧	**D,M**	SD1
2	**DTC-V2**		**Volcanic ash clouds and deposition** 🟧	**D,M**	SD2
3	**DTC-V3**		**Lava flows** 🟧	**D,M**	SD1, SD3
4	DTC-V4		Volcanic gas dispersal and deposition		SD3
5	**DTC-T1**	Tsunami	**Probabilistic Tsunami Forecasting (PTF)** 🟩	**D,M**	SD4, SD5, SD6, SD7
6	DTC-E1	Earthquake	Probabilistic Seismic Hazard and Risk Assessment		SD8
7	**DTC-E2**		**Earthquake short-term forecasting** 🟧	**D,M**	SD8, SD9
8	**DTC-E3**		**Tomography and Ground Motion Models (GMM)**	**D,M**	SD8, SD9
9	DTC-E4		Fault rupture forecasting 🟧		SD9, SD10
10	**DTC-E5**		**Tomography and shaking simulation** 🟧	**D,M**	SD8, SD11
11	**DTC-E6**		**Rapid event and shaking characterization** 🟧	**D,M**	SD8
12	**DTC-A1**	Anthropogenic	**Anthropogenic Geophysical extreme forecasting** 🟧	**D,M**	SD12, SD13

🟧 Arrows: length indicate the extent to which the AI/ML is employed (0%-100%); color indicates the relevance,

🟩 (green) AI is optional, orange required within the flow.

Summary & Conclusions

The AI/ML elements used within the DTCs are localized in field sensor-derived data processing modules spatially limiting and focusing on the specific scenario of the target area.

The DTC’s so far use mostly highly proven, effective and documented ML schemes

There is human supervision on the AI modules (no AI involve in decision making)

In some cases data gaps are filled in by AI schemes that includes reliability / probability / resolution measures

Site demonstrators aimed to test and validate the capacities of the DTC’s constitute the main tools to assess any anomalous behavior of the DTC’s due to their use in different scenarios.

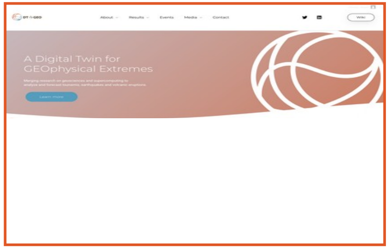
Physics based simulators

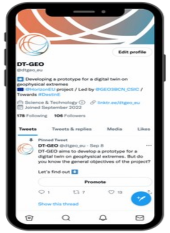
Thank you !!!

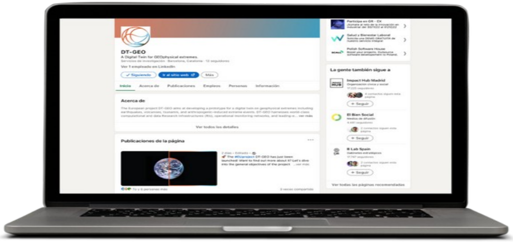
https://dtgeo.eu

@dtgeo.bsky.socialeu

linkedin.com/company/dt-geo/







Ramon.Carbonell@csic.es

This project has received funding from the European Union's Horizon research and innovation programme under the grant agreement No 101058129