E2S EMBED 2SCALE

Ref. Ares(2025)6092177 - 28/07/2025

Earth Observation & Weather Data Federation with AI Embeddings

D2.3: Benchmarking Environment for AI Compressors

The CVPR EARTHVISION Challenge and beyond

Revision: v0.2

D2.3: Benchmarking Environment for AI Compressors

Work package	WP 2
Task	T2.3
Due date	31/03/2025
Submission date	31/03/2025
Deliverable lead	Forschungszentrum Jülich (FZJ)
Version	0.1
Authors	Conrad M Albrecht (DLR), Jannik Schneider (FZJ), Stefan Kesselheim (FZJ)
()Reviewers	Isabelle Wittmann (IBM), Tim Reichelt (UOXF)
Abstract	We provide a software pipeline, implementation (Eval.AI+Python), and proof-of-concept demonstration (CVPR EARTHVISION data challenge) for a benchmark environment to test the quality of (lossy) neural compression, given a set of downstream tasks.
Keywords	lossy neural compression, CVPR EARTHVISION workshop, Eval.AI challenge platform, NeuCo-Bench framework, Earth2Vec community

Document Revision History

Version	Date	Description of change	List of contributor(s)
v0.1	Feb 20, 2025	initial submission	Conrad M Albrecht (DLR), Jannik Schneider (FZJ)
v0.2	Jul 4, 2025	revision as per feedback by Nikos Fragkoulis	Conrad M Albrecht (DLR), Jannik Schneider (FZJ), Stefan Kesselheim (FZJ)

Grant Agreement No: 101131841 | **Topic:** HORIZON-EUSPA-2022-SPACE-02-55
Call: HORIZON-EUSPA-2022-SPACE | **Type of action:** HORIZON-RIA

DISCLAIMER

Funded by the European Union
EUSPA European Union Agency for the Space Programme
Project funded by
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
Swiss Confederation
Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI
UK Research and Innovation

Embed2Scale (Earth Observation & Weather Data Federation With AI Embeddings) project is funded by the EU's Horizon Europe program under Grant Agreement number 101131841. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them. This work has also received funding from the Swiss State Secretariat for Education, Research and Innovation (SERI) and the UK Research and Innovation (UKRI).

COPYRIGHT NOTICE

Project funded by the European Commission in the Horizon Europe Programme		
Nature of the deliverable:	OTHER	
Dissemination Level	PU	
PU	Public after June 30, 2025 (past when the CVPR EARTHVISION data challenge came to a conclusion)	🌍
SEN	Sensitive, limited under the conditions of the Grant Agreement	
Classified R-UE/ EU-R	*EU RESTRICTED under the Commission Decision No2015/ 444*	
Classified C-UE/ EU-C	*EU CONFIDENTIAL under the Commission Decision No2015/ 444*	
Classified S-UE/ EU-S	*EU SECRET under the Commission Decision No2015/ 444*	

* *R: Document, report (excluding the periodic and final reports)*
DEM: Demonstrator, pilot, prototype, plan designs
DEC: Websites, patents filing, press & media actions, videos, etc.
DATA: Data sets, microdata, etc.
DMP: Data management plan
ETHICS: Deliverables related to ethics issues.
SECURITY: Deliverables related to security issues
OTHER: Software, technical diagram, algorithms, models, etc.

Table of Contents

EXECUTIVE SUMMARY

Embed2Scale (E2S) targets novel neural compression (NC) algorithms (WP2) on geospatial data (WP1) utilizing end-to-end workflows (WP3) for the benefit of real-world applications (WP4). This deliverable provides the *NeuCo-Bench* (NCB) software framework to benchmark NC that recently sparked the vibrant *Earth2Vec* (E2V) community. E2V joins stakeholders from academia, the corporate world, and governments to explore various use cases for Earth observation (EO) and Weather Data (WD).

Built on NCB, the E2S team hosted a corresponding data challenge at the 2025 IEEE Computer Vision and Pattern Recognition Conference's EARTHVISION workshop (CVPR-EVW) to engage with the relevant scientific community. As a result we teamed with the challenge winners KTH University (Sweden) & Microsoft (USA), and the University of the Bundeswehr (Germany) & the University of Boulder (USA) who scored 4th place. Corresponding algorithmic insights that advance and test the state-of-the-art in NC for SSL4EO-S12 EO data (deliverable D4.5) are presented in deliverable D2.2.

The CVPR-EVW was implemented on the well-established AI data challenge platform Eval.AI with backend at the Juelich supercomputing center in order to attract attention from the research community on the one hand, and for smooth operation on the other hand. On top, NCB and the theme of benchmarking NC for EO data has been promoted at this year's General Assembly of the European Geospatial Union (EGU25) and the European Space Agency's Living Planet Symposium (LPS25) in sessions, posters, and presentations. These activities involve the currently vibrant community revolving around Foundation Models (FMs) for EO and WD.

ABBREVIATIONS

AI	Artificial Intelligence
EO	Earth Observation
NC	Neural Compression
WD	Weather Data
CV	Computer Vision
FM	Foundation Model
Eval.AI	AI data challenge platform, https://eval.ai
E2S	Embed2Scale, https://embed2scale.eu
NCB	NeuCo-Bench, https://github.com/embed2scale/NeuCo-Bench
SSL4EO	Self-Supervised Learning for EO, dataset: https://huggingface.co/datasets/embed2scale/SSL4EO-S12-v1.1
E2V	Earth2Vec (community), slack channel https://app.slack.com/client/E27SFGS2W/C09038NSA5P (by invitation)
CVPR-EVW	Computer Vision and Pattern Recognition Conference's EARTHVISION workshop, https://www.grss-ieee.org/events/earthvision-2025
EGU25	General Assembly of the European Geospatial Union, https://www.egu25.eu
LPS25	European Space Agency's Living Planet Symposium, https://lps25.esa.int

1 Neural Compression: Assumptions & Expectations

As elaborated in great detail by the attached paper preprint [12] and as implemented in our open-source evaluation framework [13], the E2S data challenge targets:

1. **Ultra-high, Lossy Compression:** 7000 reduction rate from SSL4EO-S12 data [5] to embeddings
2. **Compression Ratio vs. Quality Trade-off:** The challenge participants developed NC encoder architectures to generate 1024-dimensional embeddings (latent vectors) to preserve semantic information as relevant for applications (SSL4EO-S12-downstream, cf. [7,8] and D4.5) such as biomass estimation, cloud characterization, land cover use fractions, and heat island statistics.
3. **Energy-Efficient Linear Probing as Quality Measure Q:** A lightweight linear model trained on-the-fly (cf. Fig. 2) generates a measure of how much semantic content is encoded by the 1024-dimensional embedding. The metric score (cf. below) builds on R-squared (regression) and F1 (classification) for image level tasks.

Q is summarized (details in [12]) as a signal-over-noise ratio quantity:

$$Q = 100\epsilon \frac{\bar{s}}{\Delta s + \epsilon} \quad \text{with} \quad \epsilon = 0.02$$

where *s* and *Δs* reflect the R-squared or F1 scores and their standard deviation on multiple training runs of the linear probing, respectively. ϵ serves as numerical regulator setting the scale at which fluctuations *Δs* turn less relevant for the computation of Q:

$$\epsilon > 0, \quad \bar{s} \in [0,1], \quad \Delta s \to 0 \quad \Rightarrow \max Q = 100$$

$$\begin{aligned} \Delta s/\epsilon \gg 1 \quad &\Rightarrow \quad q \equiv Q/100 \approx \bar{s}/\Delta s \\ \Delta s/\epsilon \approx 1 \quad &\Rightarrow \quad q \approx \bar{s}/2 \\ 0 < \Delta s/\epsilon \ll 1 \quad &\Rightarrow \quad q \approx (1 - \Delta s/\epsilon)\bar{s} \lesssim \bar{s} \end{aligned}$$

The NCB framework computes the Q-value for each downstream task. Q is designed to range from 0 to a maximum of 100 indicating the best quality of a set of embeddings. However, in practice where the fluctuations *Δs* stay finite at a minimum scale of ϵ, Q reaches maximum values of about 50. Negative Q-values indicate regression tasks with negative R-square values. In such a situation the linear probing prediction of the downstream task performs worse than a model simply predicting the mean of all task labels.

Funded by the European Union

The E2V community building around the NCB framework expects NC methods that generate Q-scores in the range of 50 when averaged over a diverse set of downstream tasks relevant to real-world geospatial applications for EO and WD.

2 The 2025 CVPR E2S Data Challenge

Scope & Community Engagement

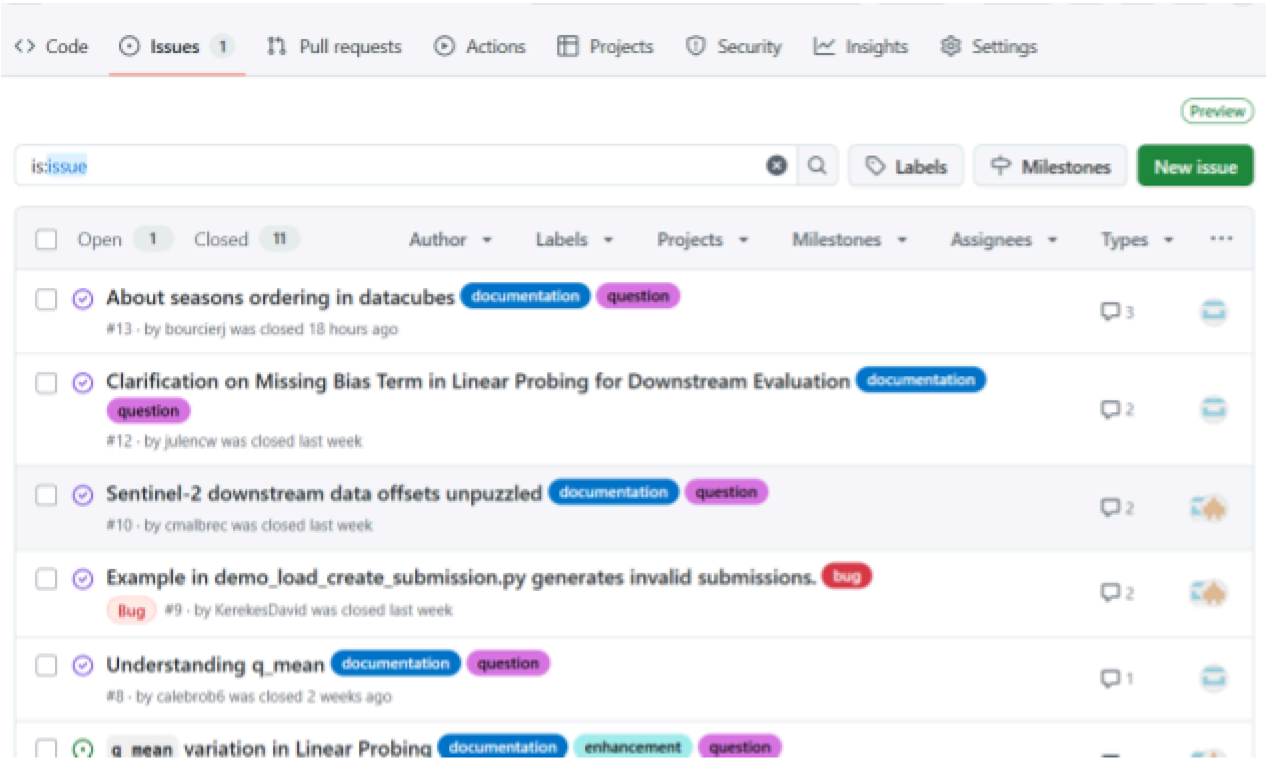

Figure 1: Snapshot of the GitHub issue section utilized to interact with the CVPR-EVW challenge participants.

The E2S data challenge [10] engaged with the EO and CV community to advance the field of NC [1]. It also served to interact with the FM community in the space of EO and WM. In a nutshell we posed the question:

If Geospatial Foundation Models claim to generate informative, generic feature vectors for a broad range of use cases, why can't we put that claim to the test?

The relevant twist to that question reads: No disclosure of the downstream tasks, but simply ask the challenge participants to embed/compress Earth observation data into 1024 floating point numbers from a given SSL4EO-S12 data cube (cf. D4.5) which amounts to a compression factor of about 7000. Additional details are provided by our paper draft [12] attached to this deliverable.

We utilized the AI data challenge platform Eval.AI as a public interface to run the challenge, and to register challenge participants. As backend running the NCB framework interacting with Eval.AI, FZJ provided a virtual machine (details below). During the challenge operation from Mar 10 through Apr 5, 2025, we built up communication through GitHub issues on a dedicated repository [11]. A snapshot of GitHub issues provides Fig. 1.

Challenge Workflow & Interfaces

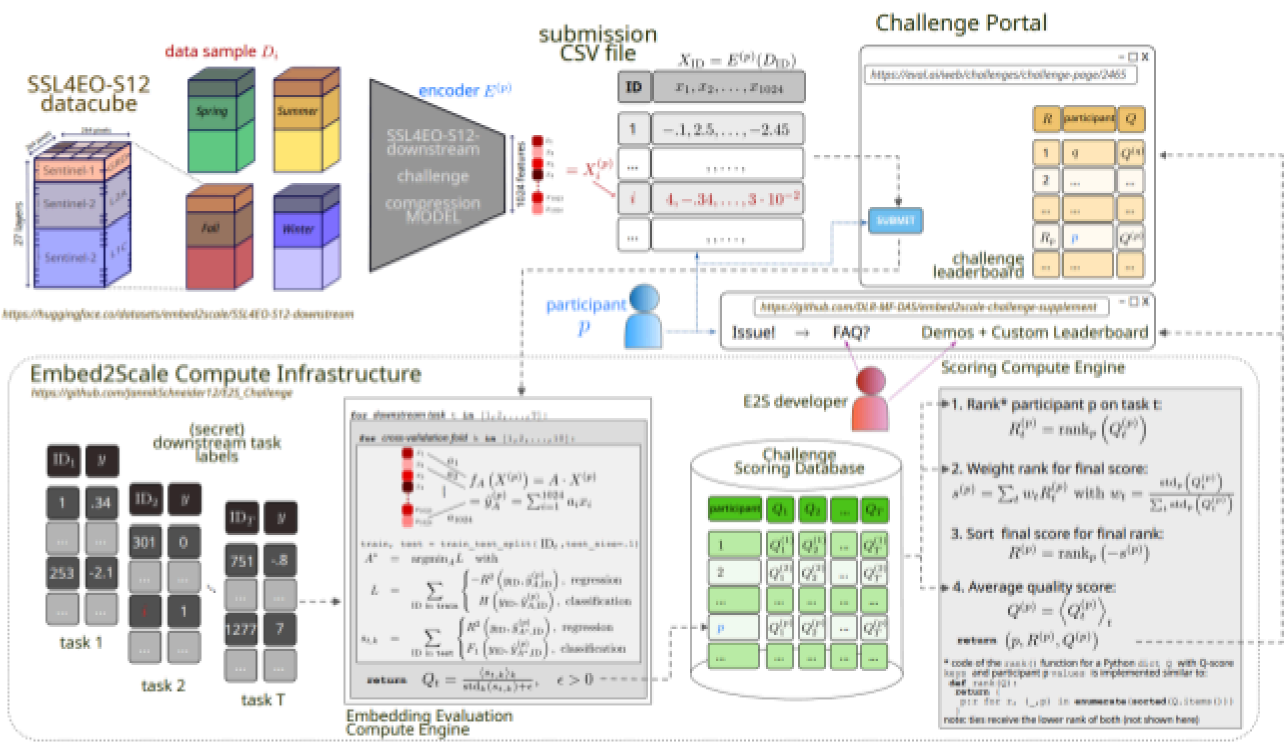

*Figure 2: **E2S data challenge overview**. The main building blocks of the challenge are the SSL4EO-S12 datacube, the submission format that the participants are expected to hand in (submission CSV file), the interface of the Challenge (Challenge Portal) as well as the setup and code for processing and evaluating the submissions and populating a leaderboard (Embed2Scale Compute Infrastructure). <u>The code resulted in the standalone NCB framework open-source software delivered hereby in [13] as seed for the E2V community currently forming.</u>*

Fig. 2 summarizes the E2S data challenge workflow, and [12] discusses the setup in length:

1. **Input Data:** The SSL4EO-S12 data cubes are specified by [3] and they are openly available through [4-6]. SSL4EO-S12 was available to the challenge participants as pretraining data without labels. The dataset which has been used for the challenge builds upon SSL4EO-S12. The participants had access to the SSL4EO-S12-downstream data without labels through HuggingFace for the *development* phase Mar 5 – Mar 31, 2025 [7] (~60 GB) and the *testing* phase Apr 1 – 5, 2025 [8] (~90GB). Embeddings were submitted in the form of CSV files where a unique, anonymous ID served as a link between SSL4EO-S12-downstream data cubes and the hidden labels utilized by the evaluation engine.
2. **Evaluation Engine:** The backend code of the E2S data challenge [9] links the Eval.AI platform with the quality measure Q. It ran on a virtual machine at Juelich Supercomputing Center. The NCB software [13] is the result of a stand-alone framework that we open-sourced based on [9]. It operates on CSV files with embeddings to generate a JSON-based leaderboard from a file-based database. NCB is lightweight to run the linear probing evaluation of a submission on commodity hardware in seconds to minutes

depending on the evaluation specific settings (number of re-evaluation for linear probing for statistics, number of epochs for training, etc.). For the E2S data challenge, the linear

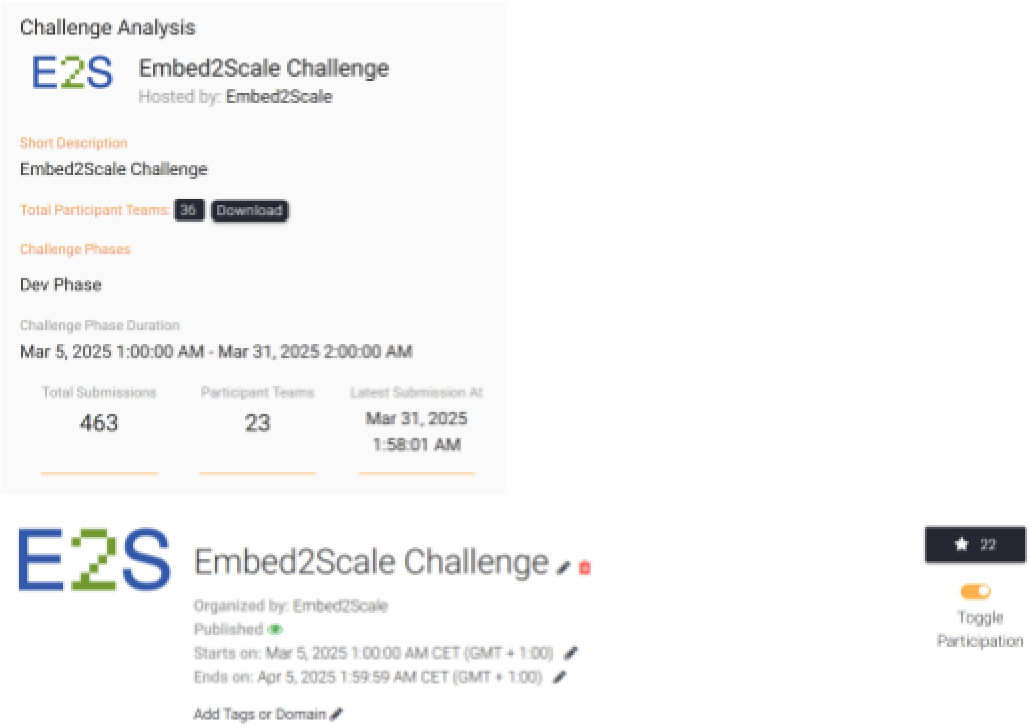

Figure 3: Screenshots of the CVPR-EVW data challenge statistics as accumulated by the Eval.AI platform for the development phase from Mar 5 through Mar 31, 2025.
Left: number of participating teams (23) and total number of development phase submissions (463). Bottom: Data challenge star ratings / "community likes" (22)

probing code operated on the set of downstream tasks linking labels and SSL4EO-S12 data cubes by the unique ID. For every downstream task (t=1,...,T in Fig. 2), a k-folded cross validation (k=1,...,10 in Fig. 2) had been performed to determine the quality score Q from R-squared or F1 statistics *s* and *Δs*. The novel scoring method is designed to compare multiple participants over multiple tasks. The participant is initially ranked per task t based on the Q-value. The final rank is calculated from the weighted mean rank across all tasks. A mean Q over all downstream tasks had been computed by a weighted average depending on the per-downstream task ranking variation: tasks where all participants scored similar in Q were considered less challenging and received a lower weight.

3. **Web Interfaces:** Given the development of their NC encoder *E*, the challenge participants submitted the embeddings (1024 floating point numbers) through the Eval.AI portal that provided feedback in form of a leaderboard based on the quality measure Q. Moreover, Eval.AI provided detailed instructions on the challenge setup, access to the challenge data, winning policies (conference participation, open-source requirements), prize money, and regulations. For interaction with the community through Q&A and raising issues, demo code, and a custom leaderboard, we set up a supplementary GitHub repository [11].
4. **Submission Policy:** The development and testing phases had their separate SSL4EO-S12-downstream datasets. While the development phase allowed for 10 submissions per day per participant, the testing phase restricted the participating teams to two final submissions. The development phase served to engineer an encoder *E* for

embeddings given feedback in terms of the quality score Q. Over the course of three weeks the teams improved the performance of their embeddings. In contrast, the testing phase minimized the number of submissions over a couple of days to avoid leakage of downstream task specifics. The leaderboard of the testing phase determined the winning solutions.

5. **Compression Baselines:** For reference performance, we introduced two baselines. One that spatio-spectral-temporally averaged the SSL4EO-S12 datacubes into 1024 floating point numbers, and a random embedding. Both “participated” as teams “Baseline random embeddings” and “Baseline mean embeddings”. All teams bet the former, but during the development phase two teams fell short of the latter baseline.
6. **Interaction of Challenge Components:** The challenge portal Eval.AI passed submissions to a virtual machine at Juelich Supercomputing Center (FZJ) that ran the evaluation engine. The GitHub repo [9] had set up a build worker based on GitHub Actions to update Eval.AI’s leaderboard. The resulting overall rank was passed back to the leaderboards on Eval.AI and a custom GitHub page through which we interacted with the challenge participants. To run the evaluation script, a CRON job was scheduled on the FZJ virtual machine. The script fetched the CSV-file submission from Eval.AI to return the final score. Moreover, the novel ranking scheme got published by a Git commit to the supplementary repository [11] where we interacted with the E2S challenge participants in GitHub issues.

The E2S data challenge as part of the CVPR-EVW highlights the practical implementation of the NCB framework. It serves as a codebase to rerun such a challenge with additional downstream tasks any time the community identifies an opportunity at a conference or similar event.

Evaluation & Insights

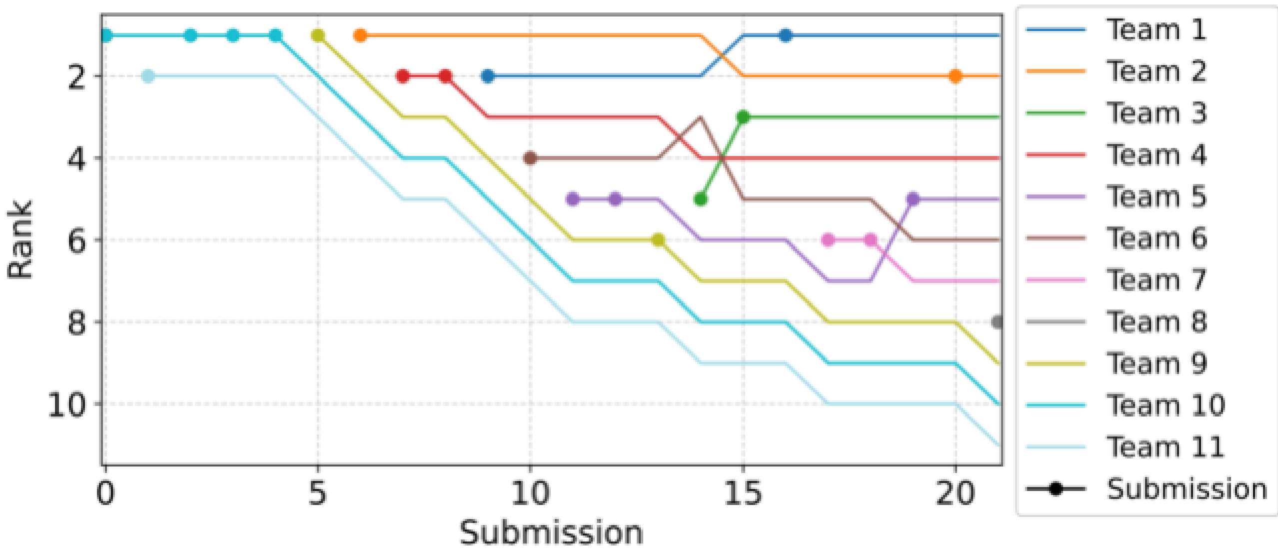

Figure 4: Dynamics of ranking for the CVPR-EVW E2S data challenge testing phase to determing the winning team.

While deliverable D2.2 will detail on the performance specifics of the winning solutions, Fig. 3 briefly summarizes statistics of the E2S data challenge as accumulated by Eval.AI. After the development phase, out of the 14 publicly visible teams (including our 2 baselines)[1], 9 participants (excluding our 2 baselines) continued to contest. Fig. 4 highlights the highly non-linear dynamics of

[1] 9 teams choose to not expose their quality score Q

the testing phase where each team was allowed just three submissions. Our ranking procedure based on the weighted quality score Q over all downstream tasks made it close to impossible for the participants to probe the hidden downstream tasks.

The quality scores Q varied widely from 0–5 on some tasks, to 5–40 on others. Our evaluation method efficiently scaled task importance ensuring that the tasks impacted the leaderboard relative to their differentiating effect across the teams. The dynamic ranking caused a swap between the original first- and second-best performing embeddings due to a third team. This highlights the impact of adaptive weighting. Additional details and analysis provides [12].

3 The NeuCo-Bench Framework for AI Compression

Referring sites

Site	Views	Unique visitors
linkedin.com	149	60
com.linkedin.android	40	20
statics.teams.cdn.office.net	20	2
github.com	11	8
lnkd.in	6	3

Popular content

Content	Views	Unique visitors
GitHub - embed2scale/NeuCo-Ben...	259	127
NeuCo-Bench/examples at main	28	13
Traffic	27	2
Pulse	23	2
NeuCo-Bench/benchmark at main	19	9
NeuCo-Bench/benchmark/evaluati...	15	6
NeuCo-Bench/examples/S2_dino_...	14	10
NeuCo-Bench/examples/data at m...	14	4
NeuCo-Bench/examples/baseline_...	13	6
add 2025 CVPR EARTHVISION dat...	13	5

Figure 5: Access statistics of the open-sourced NCB framework GitHub repository as of Jul 4, 2025.

Strategy & Community Building

As anticipated in the previous section, the CVPR-EVW E2S data challenge introduced the novel NCB framework with a strong focus on benchmarking lossy NC algorithms. Our approach builds on the promise of FMs to generate generic embeddings applicable to many downstream tasks. A detailed paper draft is provided by [12]. Our efforts aim to spark and grow an open science community around NCB to incrementally add downstream tasks, evaluation methods, and any other resources for running future competitions. While we currently focus on image-level regression and classification tasks, we understand the value of dense semantic segmentation tasks for certain applications in WP4 such as global biomass maps. Upcoming research may tap into methods such as feature-upscaling [14].

Fig. 5 provides a snapshot of the current access statistics for the NCB framework's open-source GitHub repository. We noticed spikes of access during LPS25 and the latest E2S general assembly on Jul 1-2, 2025 where advisory board members OroraTech and ICEYE had been invited.

Deliverable D2.2 presents the winning solutions of the E2S data challenge in detail:

1. **KTH University (Sweden):** Blends four Geo-FMs with a simple encoder-decoder model into a 1024-dimensional embedding space.
2. **Microsoft (USA):** Engineers four geospatial vision-language models into 1024 = 4 * 256 features.
3. **University of the Bundeswehr (Germany)**, 4th place: Exploits MOSAIKS [15] features without training to compress them down to 1024 floating point numbers by Principal Component Analysis (PCA).

In close contact with all three teams we are about to prepare a paper to summarize findings and consequences from these top-performing methods. Moreover, we intend to partner with the MOSAIKS team to add downstream tasks to deliverable D4.5. The discussions will also refine our current evaluation metrics by means of the quality score Q with potential consequences for dense tasks like pixel-wise semantic segmentation.

In fact, we recently kicked-off the E2V community at the ESA-NASA workshop on Geo-FMs [16]. As of June 2025 more than 20 organizations from industry, governments, and academia joined as summarized in Tab. 1. In monthly meetings we are going to explore the interpretation of embeddings, will discuss standards on how to integrate embeddings into APIs such as OpenEO, and we will identify future trends and directions. In an initial poll, it turned out that the Embed2Find theme is most popular among the current E2V members.

Name of Organization	**Region**	**Type**	**Weblink**
Earth Genome	US	startup	https://www.earthgenome.org
Asterisks Labs	EU	startup	https://asterisk.coop
Clark University	US	academia	https://www.clarku.edu/centers/geospatial-analytics
IBM Research Europe	EU	corporate	https://research.ibm.com
ETH Zurich	EU	academia	https://igp.ethz.ch/the-institute.html
Muenster University	EU	academia	https://cris.uni-muenster.de/portal/en/project/73726058
German Aerospace	EU	government	https://www.dlr.de/en/eoc/research-transfer/projects-missions/terrabyte
OroraTech	EU	startup	https://ororatech.com
Element 84, Inc.	US	startup	https://element84.com
Sinergise/Planet	US	corporate	https://www.planet.com
Eurac Research	EU	SME	https://www.eurac.edu
Columbia University	US	academia	https://www.eee.columbia.edu
Degas Ltd.	US	startup	https://www.degasafrica.com
LGND	US	startup	https://lgnd.io
CloudFerro	EU	startup	https://cloudferro.com
Bundeswehr University	EU	academia government	https://www.unibw.de
University of Berkeley	US	academia	https://www.ischool.berkeley.edu
Development Seed	global	startup	https://developmentseed.org
SpatialInformaticsGroup	US	startup	https://sig-gis.com
University of Zurich	EU	academia	https://dm3l.uzh.ch/en
Juelich Supercomputing	EU	government	https://www.fz-juelich.de/en/ias/jsc

Table 1: Earth2Vec community as of June, 2025

Integration with Embed2Scale & Extension for Energy Efficiency

With NCB, we provide a seed to grow an ecosystem centered around benchmarking highly-compressed embeddings on a set of standardized, community-contributed downstream tasks. To avoid contributing to an ever-growing number of benchmark datasets, we intend to harmonize with existing ones, such as MOSAIKS, GeoBench and PANGAEA. NCB's mission is to provide a standardized framework of benchmarking embeddings. Future extensions of the benchmark framework include pixel-wise and downstream tasks with temporal components (video compression).

As outlined above, the NCB framework integrates geospatial data from WP1 to provide a benchmark that evaluates the NC method development in WP2 to serve applications in WP4. Moreover, the E2V community includes stakeholders like DLR's terrabyte compute cluster, the Juelich Supercomputing Center, IBM's cloud infrastructure, CloudFerro services, and University of Muenster's openEO working group who will join forces for community standards and workflows around neural embedding APIs relevant to WP3.

Our implementation plan regarding energy benchmarking includes the ability of embeddings to be well suited for downstream model training (*Embed2Train*), because

- data and compute requirements are significantly smaller than training from scratch
- large scale analysis requires only the transfer of significantly smaller embeddings rather than the original data

For embeddings, energy is consumed in two stages:

1. the embedding generation and
2. training and inference of a downstream model

The energy consumption of generating embeddings by FMs may dominate as large geospatial datasets require large number of parameters to model correlations. In a future version of the E2S challenge, a metric involving the energy demand of data embedding generation can be taken into account. Training downstream models will typically require significantly less energy than training the embedding model as embedding quality is measured with very simple models. Only when the number of downstream tasks significantly grows, downstream model inference may turn into a dominant factor in energy consumption.

FZJ targets to implement quantification of the embedding energy consumption as follows:

1. Submission of embedding models in the form of containerized code. Eval.AI supports this use case as a *static code-upload based challenge* [17]. The submitted code will generate embeddings for previously undisclosed data.
2. In addition, these embeddings will be evaluated with one or more metrics for embedding quality, for instance the performance of a defined downstream model that is also trained during the evaluation process.
3. The evaluation code will be implemented in such a way that energy measurements based on perun software are used to evaluate the resources required for embedding generation [18].
4. The energy metric will be reported in addition to the embedding quality metric. The details of the ranking procedure need to be defined. An option: embedding models need to surpass a threshold in the metric for downstream tasks, and among those successful, energy consumption for embedding will be compared.

The above procedure aims for a fair comparison of embedding models w.r.t. energy consumption.

Funded by the European Union

EUSPA

4 REFERENCES

[1] https://www.arxiv.org/abs/2503.01505
[2] https://www.grss-ieee.org/events/earthvision-2025/?tab=challenge
[3] https://arxiv.org/abs/2503.00168
[4] https://datapub.fz-juelich.de/ssl4eo-s12
[5] https://huggingface.co/datasets/embed2scale/SSL4EO-S12-v1.1
[6] https://github.com/DLR-MF-DAS/SSL4EO-S12-v1.1
[7] https://huggingface.co/datasets/embed2scale/SSL4EO-S12-downstream/tree/main/data_dev
[8] https://huggingface.co/datasets/embed2scale/SSL4EO-S12-downstream/tree/main/data_eval
[9] https://github.com/JannikSchneider12/E2S_Challenge
[10] https://www.grss-ieee.org/events/earthvision-2025/?tab=challenge
[11] https://github.com/DLR-MF-DAS/embed2scale-challenge-supplement
[12] NeuCo-Bench: A Novel Benchmark Framework for Neural Compression (attachment)
[13] NeuCo-Bench software, https://github.com/embed2scale/NeuCo-Bench
[14] https://arxiv.org/abs/2403.10516
[15] https://www.mosaiks.org
[16] https://airdrive.eventsair.com/eventsairwesteuprod/production-nikal-public/837ec0486e4d491aa0c8f3015c87609b
[17] https://evalai.readthedocs.io/en/latest/host_challenge.html#host-static-code-upload-based-challenge
[18] http://doi.org/10.1007/978-3-031-39698-4_2