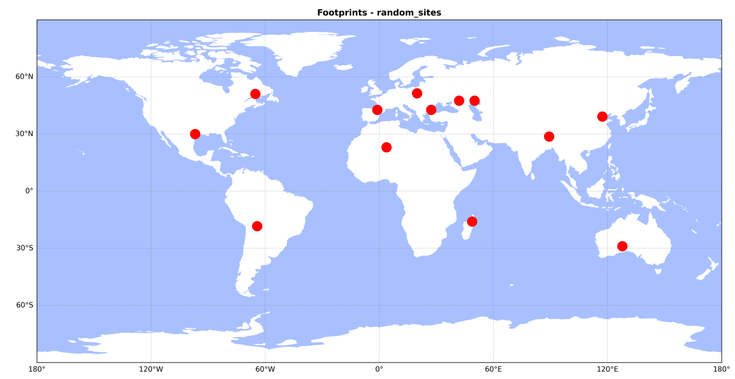
Validation of the KP Labs Multi-Class Cloud Segmentation Processor and Its Comparison to the Sentinel-2 L2A Scene Classification Map

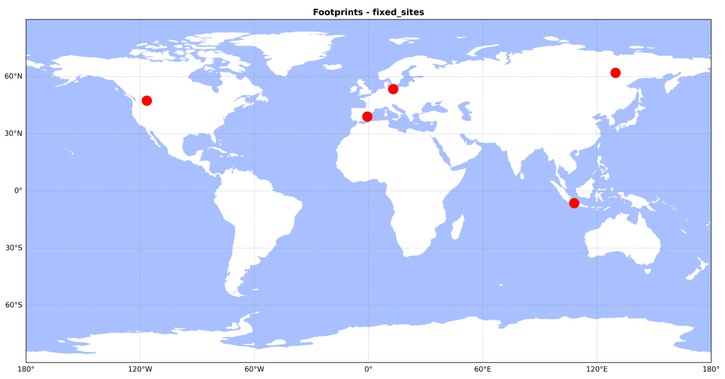
Efraín Padilla[(1)], B. Grabowski[(2)], K. Kotowski[(2)], J. Nalepa[(2,3)], R. de los Reyes[(1)], D. Marshall-Ingram[(1)], S. Clerc[(4)], S. Enache[(5)], J. Louis[(6)], F. C. Pignatale[(7)], L. De Laurentiis[(8)], V. Boccia[(8)], F. Gascon[(8)].

[1] German Aerospace Center - DLR-EOC (DE) , [2] KP Labs (PL) , [3] Silesian Univ. of Technology (PL) , [4] ACRI-ST (FR) , [5] CS Group (FR), [6] Telespazio France (FR), [7] Telespazio Germany (DE), [8] ESA-ESRIN (IT).

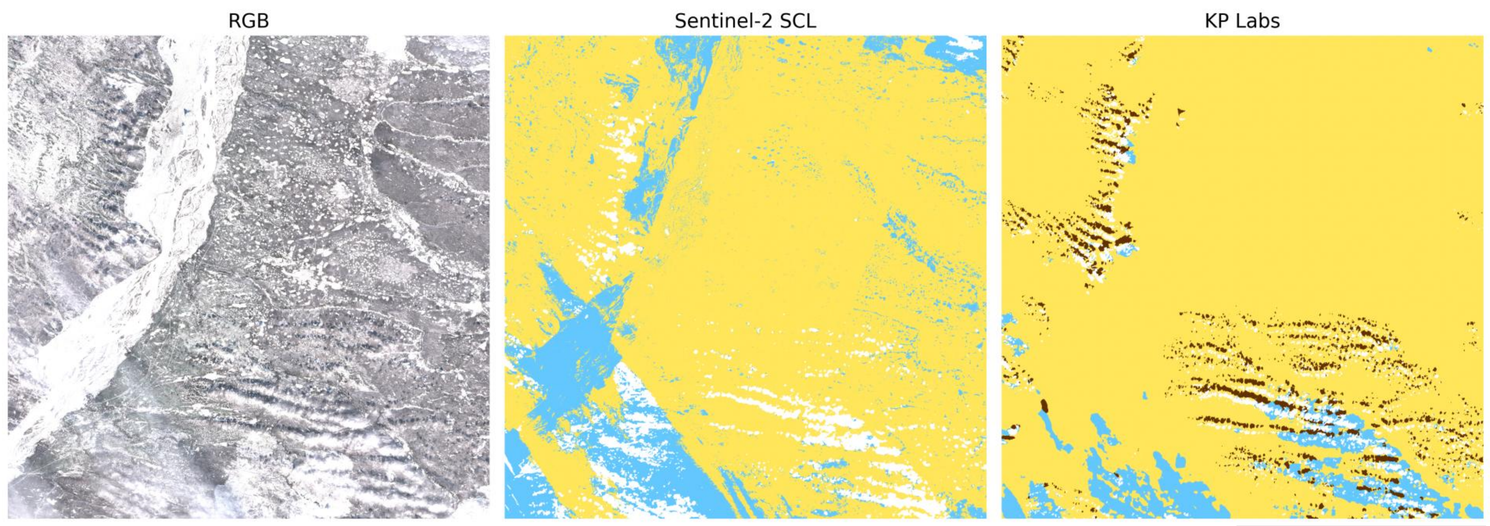
RAQRS'VIII. Torrent (Valencia), Spain. 21-25th September 2026



Spatial distribution of the 13 Sentinel-2 scenes within the random-sites dataset.



Spatial distribution of the 12 Sentinel-2 scenes within the fixed-sites dataset.



Comparison of the Yakutsk scene, Tile T52VEP (2024-03-04) within fixed-sites dataset. From left to right: Sentinel-2 quicklook, Sen2Cor classification, KP Labs classification.

Motivation

Clouds block the surface view in Sentinel-2 (S2) imagery and impacts downstream processing, so efficiently identifying cloudy images is critical for optimizing data downlinking. Accurate cloud segmentation enables "smart" on-board or on-ground data selection and compression, but machine-learning classifiers must generalize across diverse biomes and seasons. In this context, and as part of the ESA-supported AI4QC activity, KP Labs developed an AutoML-based deep-learning cloud classifier for S2.

Table 1. Mapping between Sentinel-2 SCL classes and KP Labs classification categories.

SCL Class	KP Labs Class
-Dark Area Pixels -Vegetation -Not Vegetated -Water -Unclassified -Snow	**Clear-sky**
-Cloud Shadows	**Cloud shadow**
-Cloud Medium Prob. -Cloud High Prob.	**Thick cloud**
-Thin Cirrus	**Thin cloud**
-No Data -Saturated/Defective	*Excluded*

Objective

This work provides a comprehensive validation of the KP Labs Multi-Class Cloud Segmentation Processor and compares it against the operational Sentinel-2 L2A Scene Classification Layer (SCL) produced by Sen2Cor. The goal is to quantify the classifier's accuracy, assess its ability to generalize across diverse global environments, and identify specific areas for improvement.

Methodology

The classifier was evaluated on two expert-annotated 2024 reference datasets (~677,000 pixels):

- Fixed-Sites - 12 scenes from 5 fixed locations representing different biomes, tested monthly)
- Random-Sites - 13 scenes covering every continent, spanning a wide range of land-surface and atmospheric conditions.

The KP Labs deep-learning model uses all S2 bands as input. The output classification map is 10 m GSD, but resampled to 20 m to match SCL.
The SCL classes scheme was aggregated to the 4 - cloud classes scheme for comparison (see Table 1).

Results

Quantitative performance:

The KP Labs classifier outperforms the Sen2Cor baseline on both datasets.
On Fixed-Sites dataset it improves overall nMCC by +1 p.p. (0.89 → 0.90), while on Random-Sites the gain reaches +5 p.p. (0.86 → 0.91), indicating stronger generalization across unseen biomes and atmospheric conditions. The most noticeable improvement is in the cloud shadow class, where nMCC rises by +9p.p. on Random-Sites (0.87 → 0.96), driven by a recall increase from 0.58 to 0.92 (0.61 → 0.78 on Fixed-Sites) improving the operational baseline's under-detection of shadows.

Qualitative performance:

Visual inspection shows a general reduction of noise in the classification maps, thanks to the classifier's ability to capture spatial-spectral features of the S2 imagery. Key case studies: over snow (Yakutsk), Sen2Cor misses almost all shadows due to the low spectral contrast on high-albedo surfaces, while KP Labs detects them.

Why nMCC? The Matthews Correlation Coefficient (MCC) is one of the most reliable metrics for imbalanced classification, as it only scores high when the classifier performs well in all categories of the confusion matrix, regardless of class size. It is used here instead of standard accuracy, which would be biased by the class imbalance typical in satellite imagery (e.g., clear-sky dominance). Its normalized version (nMCC) scales from 0 (total misclassification) to 1 (perfect prediction), with 0.5 equal to random guessing.

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

$$nMCC = \frac{(MCC+1)}{2}$$

Challenges:

Thin cloud remains the hardest class for both classifiers (recall ~0.4 for both on Random-Sites), due to spectral similarity and subjective class definition between thin clouds, cirrus, and clear sky.