ABSTRACT
Processing large amounts of image data such as the Sentinel-2 archive is a computationally demanding task. However, for most applications, many of the images in the archive are redundant and do not contribute to the quality of the final result. An optimization scheme is presented here that selects a subset of the Sentinel-2 archive in order to reduce the amount of processing, while retaining the quality of the resulting output. As a case study, we focused on the creation of a cloud-free composite, covering the global land mass and based on all the images acquired from January 2016 until September 2017. The total amount of available images was 2,128,556. The selection of the optimal subset was based on quicklooks, which correspond to a spatial and spectral subset of the original Sentinel-2 products and are lossy compressed. The selected subset contained 94,093 image tiles in total, reducing the amount of images to be processed to 4.42% of the full set.

1. Introduction
The Copernicus program of the European Commission (EC) with its Sentinel satellites produces approximately 10 TB of Earth Observation (EO) data per day. This wealth of information, combined with a free full and open access policy, provides new opportunities for applications in forestry, agriculture, and climate change monitoring, to name a few. An increasing number of platforms are being created that address the storage and processing of these data, both from the institutional and the private sector. The need for computing power that can process the amount of available data can only be expected to increase. The EC has launched an initiative earlier this year to develop Copernicus Data and Information Access Services (DIAS) that facilitate access to these data and allow for a scalable computing environment.

Nevertheless, computing power comes at a cost, both financially and environmentally (Whitehead, Andrews, Shah, & Maidment, 2014). A typical data center with thousands of servers may consume as much energy as 25,000 households (Dayarathna, Wen, & Fan, 2016). Processing large amounts of data also take valuable time, not only from the servers, but also from the users that have to wait for exploiting their results. Our contribution with this paper is to provide means to reduce the computations in a twofold approach. First, we
The bit encoded image defines which pixels should be considered for the final composite. For instance, pixels encoded as the integer value 7 (binary value “00111”) are within the

minimize the number of images that must be processed with little impact on the quality of the final result. Second, instead of dealing with the original images at full resolution, the selection criterion is based on a sample only. In our case, this was performed by considering the quicklooks of the images, that is a spatial and spectral subset of the original images.

The reduction of the number of images is driven by the fact that for most applications, not all acquired images are equally valid. Often, the useful information can be obtained from a selection of the images. The challenge that was tackled in this paper was to find the minimum set of images that maximizes the information content for the problem at hand. A fitness criterion was therefore defined that evaluates each image. The underlying idea is that the computational cost of the optimization function is considerably less than the processing algorithm that is needed to produce the application results.

This paper is organized as follows: In Section 2, we present the materials used for the chosen application as well as the infrastructure on which the tests were implemented. In Section 3, the methods are presented, and in Section 4, results are shown. The conclusions are drawn in Section 5.

2. Materials and infrastructure

As an illustration of the optimized image selection in a Big Data context, we focused on the application of the creation of a global cloud-free satellite image composite of the land surface. This application can serve other derived products, for instance the identification of human settlements (Pesaresi et al., 2016). The selection was made from 2,128,556 optical remote sensing images. This corresponds to the full set of images acquired with the Sentinel-2A sensor between January 2016 and September 2017 (see Figure 1). Sentinel-2A and Sentinel-2B are a constellation of two identical optical sensors with 13 spectral bands. Together, they almost cover the entire land surface of the globe every five days. The original images have a spatial resolution of 10 to 60 m, depending on the spectral band. They daily generate about 1.6 TB of compressed raw image data (Drusch et al., 2012). The European Space Agency (ESA) provides the Sentinel-2 images in 100 × 100 km tiles according to the Military Grid Reference System (MGRS). A total of 29,782 distinct MGRS tiles were identified for this application that cover the acquired Sentinel images.

In addition to the original images, ESA delivers true color quicklooks at a reduced spatial resolution of 320 m, as well as a cloud mask in vector format (see Figure 2). The combination of the data domain and cloud mask was bit encoded according to the encoding scheme described in Table 1. Pixels in the quicklook with a 0 value were considered to be outside the data domain.

The bit encoded image defines which pixels should be considered for the final composite.
Figure 1. Images acquired between January 2016 and September 2017 with the Sentinel-2A sensor: daily and cumulative count.

Figure 2. Sentinel-2 quicklook (a) and mask defining cloud and data domain (b) from MGRS tile 31UES, relative orbit 94, 23 June 2016.

data domain for all three spectral bands and have not been detected as cloudy (neither opaque clouds nor Cirrus). Those pixels are candidates for the global image composite. In Figure 2(b), they are represented in gray. Clouds are represented in blue, whereas pixels not within the data domain are shown in white. Notice that not all clouds have been detected,
which is a current limitation of the delivered cloud mask. Due to compression artifacts of the quicklook images at the high-contrast edges near the border of the data domain, we performed a morphological opening on the data domain.

The infrastructure that is available in house, referred to as the Joint Earth Observation Processing Platform (JEODPP) for this analysis is based on commodity hardware and open source software (Soille, Burger, Rodriguez, Syrris, & Vasilev, 2016; Soille et al., 2017a, 2017b). The Linux operating system is run on both the storage and processing nodes. There are 39 processing nodes with 952 cores and 15.3 TB of RAM in total. The full storage, currently at 1.9 PB, is served through a distributed file system (DFS). The implementation of the DFS is EOS (Mascetti, Gonzalez Labrador, Lamanna, Moscicki, & Peters, 2015), which has been developed by the European Organization for Nuclear Research (CERN). For the workload manager, HTCondor (Thain, Tannenbaum, & Livny, 2005) was selected. It is particularly fit for high-throughput computing, we typically deal with when processing Earth observation data.

The programs for the image selection are written in Python code, but the underlying image processing functions are written in C and C++ (McInerney, & Kempeneers, 2014; Soille, 2004). These functions have been made available in Python via the Simple Wrapper Interface Generator (SWIG). The input/output (I/O) functions accessing geospatial data heavily rely on the Geospatial Data Abstraction Library (GDAL).

3. Methods

Due to cloud cover that occurs in different parts of the globe, multiple acquisitions acquired at different dates are required for a global cloud-free composite. For some regions, images acquired at a single acquisition date in cloud-free conditions can be found. For other regions, due to persistent cloud cover, overlapping images acquired at different times must be combined in order to obtain a cloud-free composite. For instance, in equatorial South America, the Congo River basin in Africa, and Southeast Asia, the probability of cloudy observation clouds is typically more than 80% (Wilson, & Jetz, 2016). The objective here is to find the minimum set of overlapping images from which a cloud-free global composite can be created. The image selection is performed in two steps: a first selection based on a quality check and the actual optimization scheme that finds the minimum set from this reduced set.

3.1. Quality check

The full set of 2,128,556 images acquired with the Sentinel-2A sensor between January 2016 and September 2017 was first reduced according to some quality checks. All images with a Sun zenith angle greater than 70 degrees were discarded. Acquisitions obtained with large Sun zenith angles increase the number of shadows in the image, which have a negative impact on the image quality. Especially in the higher latitudes, the reduction of images due to this filter was important.

A second reduction was based on the average pixel value in the blue band, calculated on the cloud-free data domain of the image tile. This value was used as a proxy for the level of contamination due to atmospheric effects. Indeed, the reflectance in blue due to path radiance is typically much higher than for the green and red bands (Kaufman et al.,
Therefore, all overlapping image tiles for an MGRS tile were ordered according to their average pixel value in blue. Only images corresponding to the lower 50 percentile were retained. If the reduced set was not sufficient to create a cloud-free composite for the respective MGRS tile, the percentile was incrementally increased. Notice that each MGRS tile can be visited from multiple relative orbits, with a potentially different coverage of the MGRS tile. Therefore, the percentiles were calculated for each relative orbit separately.

Finally, some quality checks were introduced to remove inconsistent image tiles in the full set. For instance, duplicate images were found with different production dates (only the latest product was retained). A number of image tiles with inconsistent MGRS tile names was also excluded.

### 3.2. Optimization scheme

The 290 km wide swath of any given relative orbit is only partially covering the MGRS tiles intersecting the boundary of the swath. We therefore created a binary mask that corresponds to the data domain, i.e. pixels where valid data are present as explained in Section 2.

Consider the synthetic example of three overlapping image tiles, for which the corresponding binary masks are shown in Figure 3. The optimization consists of selecting the minimum set of images that covers the data domain without cloud contamination. Starting from the least cloudy image 1, three images (a, b, and c) are required to obtain a cloud-free composite. This example shows how we can fall into a local minimum that results in a suboptimal set. Indeed, the minimum set is composed of only two images (b and c). A global optimal set can be selected using an exhaustive search through all the combination of all $N$ available images. However, the number of combinations grows exponentially with the number of available images ($2^N$). We used the sequential floating forward search algorithm (SFFS) of Pudil, Novovičová, and Kittler (1994), which avoids the local minimum and is able to find the correct set of images (b and c) in the example of Figure 3.

**Algorithm 1** Sequential Floating Forward Search of minimum image set

*Inputs:* overlapping image set $Y = \{y_j | j = 1, \ldots, N\}$ and maximum number of selected images $M_{\text{max}}$, where $1 \leq M_{\text{max}} \leq N$

*Output:* selected images $X = \{x_j | j = 1, \ldots, M\}$, where $x_j \in Y$ and $1 \leq M \leq M_{\text{max}}$

*Initialization:* $X_0 = \emptyset$; $k = 0$

repeat

{Inclusion: add the most significant image with respect to $X_k$}

$x_j^+ = \arg\max_{x_j \in Y \setminus X_k} J(X_k \cup x_j)$

$X_{k+1} = X_k \cup x_j^+; \quad k = k + 1$

repeat

{Conditional exclusion: try to remove images}

$x_j^- = \arg\max_{x_j \in X_k} J(X_k \setminus x_j)$ [the least significant image in $X_k$]

$X_{k-1} = X_k \setminus j^-; \quad k = k - 1$

until $J(X_k \setminus j^-) < J(X_{k-1})$

until $k = M_{\text{max}}$

$X = X_k$
Figure 3. Cloud-free data domains (colored area) for three overlapping images: mask1 (a: 15% no data), mask2 (b: 29% no data) and mask3 (c: 29% no data).

Table 2. Distribution in percentage of number of overlapping image tiles ($M$) in the optimal subset.

| $M$ | 1   | 2   | 3   | 4   | 5   |
|-----|-----|-----|-----|-----|-----|
|     | 26% | 17% | 12% | 7%  | 38% |

The iterative algorithm is listed in Algorithm 1. It selects the minimum set of $M$ images ($X_M$) from the set of overlapping images $Y$ for a specific tile to obtain a cloud-free composite. We start with an empty set $X_0$. At each iteration, we try to include the most significant image with respect to $X_k$, with $k$ the current number of images in the selected set. The objective function $J(X_k)$ is based on the number of cloud-free pixels within the data domain. To avoid a local minimum, we try to remove the least significant image in $X_k$. If the objective function of the reduced set is lower than that of the set of the previous iteration ($X_{k-1}$), we replace the current set with the reduced set. In the original algorithm proposed (Pudil et al., 1994), the iteration continuous until the set $X_k$ reaches the required dimension. Here, we slightly adapted the algorithm and stop the iteration when a cloud-free composite is obtained since in this case the objective function cannot be further improved. If a single cloud-free image covers the entire tile, the number of images equals one ($M = 1$). Notice that some pixels that cover highly reflective areas on the ground can result in persistent commission errors of the cloud detection algorithm. We therefore constrained the dimension of the set to a maximum value ($k \leq M_{\text{max}}$). Without this constraint, the algorithm would select the entire overlapping image set ($M = N$), in an attempt to obtain a cloud-free composite.

4. Results

Results are presented in two parts. In Section 4.1, we focus on the image selection process. Results show both the number of selected image tiles as well as how they are geographically distributed. In Section 4.2, the processing performance is reported. We also present results on the scalability of the selection task using the JEODPP cluster.

4.1. Image tile selection

The full set of 2,128,556 images was reduced to 1,071,123 images after the quality check described in Section 3.1. This corresponds to an average of 36 overlapping images per MGRS tile. The maximum overlap was 867 for a tile located in Greenland (tile T25XEJ at a
Figure 4. Geographical distribution of the reduced image set per MGRS tile, after the quality check.
Figure 5. Number of overlapping image tiles ($M$) in the optimal subset obtained from the selection algorithm with $M_{\text{max}} = 5$. 
Figure 6. Percentage of valid data in the image composites (cloud free and within data domain).
Figure 7. Selected set of five overlapping images (a–e) and resulting minimum composite in blue (f) for MGRS tile T20NMG, Amazonian (Brazil), with persistent cloud cover (pixels in white).

latitude of 80 degrees North) that is covered by 41 relative orbits. A geographic distribution of the overlapping images in the filtered set is shown in Figure 4.

From the reduced set, a total of 94,093 images were then selected with the proposed optimization scheme. These images correspond to the minimum set from which a global cloud-free composite can be created. The number of overlapping image tiles required for a cloud-free composite \( M \) in each MGRS tile is shown in Figure 5.

The average number of \( M \) was 3.16. For 26% of all MGRS tiles, a single image was selected \( (M = 1) \). The remaining MGRS tiles required a combination of overlapping images to obtain a cloud-free composite (see also Table 2). At least five image tiles were required for 38% of the MGRS tiles.

In Figure 6, the percentage of cloud-free pixels within the data domain is shown for the optimal subset in each MGRS tile. Persistent cloud cover were an issue in mountainous areas such the Himalaya and the Andes, but also in Antarctica (indicated in orange and red). However, some of the cloud detected areas are in fact matching snow-covered areas. Another example of persistent cloud cover is shown in Figure 7 for an image in the tropical forest. All five selected images (Figures 7(a–e)) are contaminated with clouds. Also the resulting composite image, shown in Figure 7(f), has some persistent cloud cover. Although the actual composition is not the main topic here, shadow pixels are an issue. Due to the minimum blue composite rule, shadow pixels are likely to be selected. We therefore
Figure 8. World composite based on selected Sentinel-2 quicklooks.
adapted the minimum composite rule by first calculating the morphological dilation of the blue band.

4.2. Processing performance and scalability

The selection process was performed on the JEODPP cluster on an MGRS tile by tile basis. With no dependency or need for communication between the tasks, this is a typical example of an embarrassingly parallel workload. Using 952 cores, the total processing time on the JEODPP cluster was 20 hours. The composite step of the selected quicklooks, based on the minimum pixel value in the blue band, was included in this process. The result is a true color global land composite at a spatial resolution of 320 m, shown in Figure 8. Most of the MGRS tiles were effectively composited as cloud free.

The scalability of the selection task was tested with a fixed set of 912 MGRS tiles that was processed five times. The number of cores was increased each time from 64 to 128, 256, 512, and finally 912. These five jobs processed as many MGRS tiles in parallel as the number of cores available. They were run through a directed acyclic graph (dag) manager with HTCondor, designed to run the jobs in sequence so that a job was only started once the previous job was completed. This allowed for a fair comparison between the processing times of each job, which had all resources (e.g. network and memory) available. The CPU usage for the consecutive jobs is shown in Figure 9(a). As expected, it increases when more cores are available, whereas the processing time is decreased. The throughput in MGRS tiles per hour is shown in Figure 9(b), suggesting a linear scalability up to 912 cores. The error bars shown are calculated as 26.34 MGRS tiles per hour, based on the root mean square error of the linear regression \( R^2 = 0.99 \).

5. Conclusion and future outlook

An optimization scheme was presented that selects a subset of the Sentinel-2 archive in order to reduce the amount of processing, while retaining the quality of the resulting...
output. As a case study, we focused on the selection of a cloud-free composite, covering the global land mass and based on the images acquired from January 2016 until September 2017. The selection of the optimal subset was based on quicklooks, which correspond to a spatial and spectral subset of the original Sentinel-2 products and are lossy compressed. Using the optimal subset instead of the full set leads to an important reduction in processing time and resources. The 94,093 image tiles in the optimal subset represent only 4.4% of the 2,128,556 images acquired in the period January 2016 until September 2017. The average number of image tiles in the full set was 72 per MGRS tile, which was reduced to 3.16 image tiles per MGRS tile for the optimal subset. For the selection, all image tiles had to be processed, but only at a spatial resolution of 320 m. This corresponds to 0.01% of the original data volume at 10 m resolution.

In this case study, we focused on a global land cloud-free composite. The next step would be to use the optimal subset to create such a composite at full spatial resolution. For instance, using the visual spectral bands in red, green, and blue of the Sentinel-2 sensor, a true color composite can be produced at a spatial resolution of 10 m. A number of issues need to be addressed. Due to restrictions of the available cloud mask, undetected clouds remain in the final composite. This problem was mitigated by introducing a quality check, based on the average pixel value in the blue band prior to the optimal subset selection. In addition, we used a composite rule that was based on the minimum pixel value in the blue band. Although most of the undetected clouds could be removed in this way, another issue related to shadow pixels was emphasized. A better cloud mask as well as the introduction of a new mask indicating cloud shadows could greatly improve the results. Finally, the optimal subset selection algorithm can be easily adapted to application driven criteria by modifying the rules for the quality check and those at the basis of the objective function.

Data availability statement

The Sentinel-2A quicklooks are available via the Copernicus Open Access Hub (https://scihub.copernicus.eu/).

Disclosure statement

No potential conflict of interest was reported by the authors.

References

Dayarathna, M., Wen, Y., & Fan, R. (2016). Data center energy consumption modeling: A survey. *IEEE Communications Surveys & Tutorials, 18*(1), 732–794.

Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., ... Meygret, A. (2012). Sentinel-2: ESA’s optical high-resolution mission for GMES operational services. *Remote Sensing of Environment, 120*, 25–36.

Kaufman, Y. J., Wald, A. E., Remer, L. A., Gao, B.-C., Li, R.-R., & Flynn, L. (1997). The MODIS 2.1 μm channel-correlation with visible reflectance for use in remote sensing of aerosol. *IEEE Transactions on Geoscience and Remote Sensing, 35*(5), 1286–1298.

Mascetti, L., Gonzalez Labrador, H., Lamanna, M., Moscicki, J. T., & Peters, A. J. (2015). CERNBox + EOS: end-user storage for science. *Journal of Physics: Conference Series, 664*(6), 062037.

McInerney, D., & Kempeneers, P. (2014). Chap. 12. Pktools. In B. Bernd, W. Lahoz, & D. P. Solomatine (Eds.), *Open source geospatial tools, earth systems data and models* (pp. 173–197). Cham: Springer-Verlag.
Pesaresi, M., Corbane, C., Julea, A., Florczyk, A. J., Syrris, V., & Soille, P. (2016). Assessment of the added-value of Sentinel-2 for detecting built-up areas. Remote Sensing, 8(4), 299.

Pudil, P., Novovičová, J., & Kittler, J. (1994). Floating search methods in feature selection. Pattern Recognition Letters, 15(11), 1119–1125.

Soille, P. (2004). Morphological image analysis: Principles and applications (Corrected 2nd printing of the 2nd ed.). Berlin: Springer-Verlag.

Soille, P., Burger, A., De Marchi, D., Hasenohr, P., Kempeneers, P., Rodriguez, D., ... Vasilev, V. (2017a, November 27–30). The JRC earth observation data and processing platform. Proceedings of the Conference on Big Data from Space (BiDS’17), Toulouse, France.

Soille, P., Burger, A., De Marchi, D., Kempeneers, P., Rodriguez, D., Syrris, V., & Vasilev, V. (2017b). A versatile data-intensive computing platform for information retrieval from big geospatial data. Future Generation Computer Systems, 81, 30–40. doi:10.1016/j.future.2017.11.007

Soille, P., Burger, A., Rodriguez, D., Syrris, V., & Vasilev, V. (2016). Towards a JRC earth observation data and processing platform. Proceedings of the Conference on Big Data from Space (BiDS’16) (pp. 15–17). Spain: Santa Cruz de Tenerife.

Thain, D., Tannenbaum, T., & Livny, M. (2005). Distributed computing in practice: The condor experience. Concurrency and Computation: Practice and Experience, 17, 2–4.

Whitehead, B., Andrews, D., Shah, A., & Maidment, G. (2014). Assessing the environmental impact of data centres part 1: Background, energy use and metrics. Building and Environment, 82, 151–159.

Wilson, A. M., & Jetz, W. (2016). Remotely sensed high-resolution global cloud dynamics for predicting ecosystem and biodiversity distributions. PLoS Biology, 14(3), e1002415.