Snow and cloud classification in historical SPOT images: An image emulation approach for training a deep learning model without reference data

Zacharie Barrou Dumont, Simon Gascoin, Jordi Inglada

Snow and cloud classification in historical SPOT images: An image emulation approach for training a deep learning model without reference data

Journal:	*Journal of Selected Topics in Applied Earth Observations and Remote Sensing*
Manuscript ID	JSTARS-2023-00896
Manuscript type:	Special Section: AI Empowered Monitoring of Climate and Environmental Changes by Remote Sensing
Date Submitted by the Author:	08-Jun-2023
Complete List of Authors:	barrou dumont, zacharie; Paul Sabatier University, CESBIO Gascoin, Simon; UPS/CNES/CNRS/IRD, CESBIO Inglada, Jordi; Centre d'etudes spatiales de la biosphere, CESBIO
Subject Category Please select the subject category that best reflects the scope of your manuscript:	Methodologies and Applications to: Cryosphere
Keywords:	Snow, Remote sensing, Artificial intelligence, Neural networks, Clouds, Time series, Image classification

1

Snow and cloud classification in historical SPOT images: An image emulation approach for training a deep learning model without reference data

Zacharie Barrou Dumont, Simon Gascoin, and Jordi Inglada

***Abstract*—The lack of revisit in long-term satellite time series such as Landsat is an issue to assess ecosystems response to snow cover variations in mountains. A recent release of the SPOT 1-5 satellite images collection by the SPOT World Heritage (SWH) program offers the opportunity to increase the temporal revisit of Landsat from 1986 to 2015 at 20 m resolution. However, spectral and radiometric limitations of these images hinder the application of well-established pixel-wise methods to extract the snow cover area. As a work-around, deep learning techniques such as convolutional neural networks can incorporate both spectral and spatial information to classify every pixel as snow, cloud, or snow-free. However, the lack of reference data poses a challenge to the implementation of such data-driven approaches. Here, we develop an emulator of SPOT images which takes as input Sentinel-2 images. As a result, an emulated SPOT image can be paired with a reference snow map generated from its source Sentinel-2 image to train a deep learning model able to process actual SPOT images. We follow this approach to train a U-Net and evaluate different training strategies. We apply the different models to classify actual SPOT images for which we have reference data for validation. The method yields high precision in detecting snow, with minimal false snow pixel identification. This is at the cost of overestimating cloud pixels around clouds and highly saturated areas. The results confirm the potential of this method to generate time series of snow cover maps using the SWH collection.**

***Index Terms*—snow cover, image classification, SPOT World Heritage, Sentinel-2, U-Net, deep learning.**

I. Introduction

THE snow cover area (SCA), defined as the spatial extent of the snow cover on the land surface [1], is an important variable to understand hydrological and ecological processes in mountainous regions. In particular, mountain soil and vegetation properties are largely driven by seasonal snow cover duration and snow melt-out date [2]–[6]. Therefore, information on the spatial-temporal variability of the snow cover area over long time periods is critical to study the response of mountain ecosystems to climate change. In addition, high resolution observations are needed since the mountain snow cover varies at spatial scales typically of 100 m or below due to topographic heterogeneity [7].

The Landsat program provides the opportunity to map the extent of snow cover at decametric resolutions (60-30 m) since 1972 with theoretical revisit times of 16 to 8 days [8]. However, the theoretical revisit time is not always guaranteed due to technical obstacles, mission constraints and cloud cover [9], [10]. Considering that the cloud cover probability often exceeds 50% in temperate mountainous regions [11], [12], such revisit time enables approximately one observation per month or less, which is insufficient to characterize the seasonal evolution of the snow cover [13].

The United States Geological Survey also provides free access to the data acquired by the Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) instrument on the Terra satellite, which has been capturing images since 2000 with resolutions of 15 to 90 meters. The ASTER instrument does not, however, allow systematic acquisitions and only collects an average of 8 minutes of data per 99-minute Sun-synchronous orbit, i.e. 650 scenes per day [14].

The French Space Agency (Centre National d'Études Spatiales, CNES) led the SPOT (Satellites Pour l'Observation de la Terre) program (https://spot.cnes.fr/) by launching five Earth observation satellites between 1986 and 2002. The SPOT satellites observed the Earth in the visible and infrared bands with spatial resolutions of 10 and 20 meters. Accessible for free since 2015, nearly 20 millions of SPOT 1 to 5 products from acquisitions between 1986 and 2015 can be obtained through the SPOT World Heritage (SWH) program at the 1A processing level (without orthorectification) [15].

Because SPOT image acquisitions were performed on a demand basis rather than systematically, it is challenging to determine the average revisit frequency. However, SWH can significantly increase the number of available observations. For example, SPOT products account for approximately half of the acquisition dates over the French Alps and Pyrenees that could be obtained from Landsat and ASTER only over 1996-2005 [16].

Although higher level products or processing algorithms are readily available to extract snow cover maps from Landsat and Sentinel-2 missions [17], [18], there is no equivalent for SWH. Well-established methods for snow/cloud discrimination on Landsat and Sentinel-2 images primarily rely on pixel-based spectral signatures such as the Normalized Snow Difference Index (NDSI) [18], [19] or spectral unmixing [20]. Both approaches take advantage of the specific spectral properties of snow cover surfaces, highly reflective in the visible but not in the shortwave infrared wavelengths (SWIR, typically 1.6 μm), whereas clouds tend to have a higher reflectance in the SWIR [21]. SWH data are encoded in 8-bits, with only 256 different reflectance values per band, and often saturate when clouds or snow are present. This saturation, combined with the

Centre d'Etudes Spatiales de la Biosphère (CESBIO), Université de Toulouse, CNRS/CNES/IRD/INRAE/UPS, 18 avenue E. Belin, 31401 Toulouse.

TABLE I
SPATIAL RESOLUTIONS OF SATELLITE MISSIONS

Mission	Spatial resolution (m)		Period
	Visible	SWIR	
Sentinel-2	10	20	2015-present
SPOT1-3	20	-	1986-2009
SPOT4	20	20	1998-2013
SPOT5	10	20	2002-2015

lack of SWIR band in SPOT 1-3 products, makes it difficult to differentiate snow from clouds with a per-pixel approach. A solution to overcome these issues is to use deep learning approaches which can take into account both the spectral and spatial information to segment a satellite image into snow, clouds and ground [22], [23].

Despite the slow uptake of deep learning methods in the remote sensing field, there has been a rapid increase in the number of studies utilizing these techniques [24]. Deep learning has been used for various remote sensing tasks, such as road detection, sea-land detection, land cover mapping and cloud detection. In particular, Convolutional Neural Networks (CNNs) have emerged as a highly effective tool in remote sensing image classification (semantic segmentation in deep learning terminology). A significant advantage of CNNs over previous methods is their ability to automatically extract features, which used to be a manual task. Specifically, a CNN architecture developed for biomedical image segmentation, U-Net [25] was used to develop a cloud detection algorithm for Landsat 8 imagery trained on annotations from the Fmask processor and has shown increased performances compared to the processor itself [26]. However, the lack of reference snow cover maps corresponding to SWH images impedes the implementation of such approaches.

In this work, we present a novel approach for training a U-Net to generate thousands of snow cover maps from SPOT images at 20 m resolution. We developed an emulator of SPOT images which takes as input a Sentinel-2 image. As a result, the emulated SPOT image (pseudo-SPOT) can be paired with a reference snow map generated from its source Sentinel-2 image. We used this approach to train the U-Net over the Pyrenees. The trained U-Net can then be used to segment actual SPOT images into a snow, clouds and ground map. The advantage of this method is twofold: the ability to do snow detection with highly-saturated images without a SWIR band, making it most useful when processing SPOT 1-3 images, and the ability to train and use a U-Net model over any region covered by SWH images thanks to the global coverage of Sentinel-2. The method is evaluated by testing the U-Net's performances in learning from pseudo-SPOT4 images and inferring on real SPOT4 images. SPOT4 images have a SWIR band which we can use to generate reference maps using a pixel-based approach.

II. DATA

A. SPOT

Each SPOT had two identical instruments. The first generation SPOT1 to 3 were equipped with twin High Resolution in the Visible (HRV) instruments with green, red and near-infrared bands at 20 meter spatial resolution. SPOT4 was equipped with identical instruments (HRV-IR) excepted for an additional short-wave-infrared band (SWIR) at the same resolution. The SPOT5 High Geometrical Resolution (HRG) twin instruments had the same characteristics as HRV-IR excepted for an improved spatial resolution of ten meters for the three visible bands (Table I). The swath of SPOT multispectral sensors was 60 km.

SPOT pixel values were encoded in 8 bits (only 256 different values). The conversion from the instrument outputs to radiance values was done via a ground controlled preset gain chosen depending on the context of the image acquisition to optimize the image dynamics. This radiometric gain was generally adjusted for vegetation surfaces and therefore often saturates over snow and clouds.

We used SPOT images at the 1C processing level (orthorectified and top-of-atmosphere reflectance) available from the Theia catalog (https://www.theia-land.fr/product/spot-world-heritage-fr/). The images are distributed as integers in milli-reflectances (reflectance x $1O^3$) in 16-bit unsigned integer but actually contain only 256 values.

B. Sentinel-2

The Sentinel-2 mission is a constellation of two satellites (2A and 2B). Both Sentinel-2 A and B use the same Multi Specral Imager (MSI) with thirteen bands in the visible and infrared ranges and with spatial resolutions of 10, 20 and 60 meters [27]. Following the launch of Sentinel-2B in 2017, the Sentinel-2 mission observes the global land surface with a revisit time of 5 days (Table I).

In this work, we use Sentinel-2 products at level-1C to match the processing level of the SWH data. Sentinel-2 level-1C imagery is projected on a grid of 100x100km2 tiles defined by the Military Grid Reference System (MGRS). MSI sensitivity allows precise radiometric measurements over a large range of reflectance and therefore saturated pixels are very rare even over bright clouds or snow covered surfaces.

We also use level-2B products from Theia, providing a classification of the land surface at 20 m resolution into four classes: snow, no snow, cloud (including cloud shadow) and no data. The Theia snow products are generated using the MAJA and LIS software [18]. MAJA performs the atmospheric correction and cloud detection to generate level-2A products [28], which are then used as input to LIS to perform the snow detection. The performance of the snow detection using the MAJA-LIS pipeline was assessed on the French Alps and Pyrenees with an accuracy of 94% (kappa 0.83) [18]. A more comprehensive evaluation was conducted at pan European scale and yielded comparable results [29]. However, where transparent or semi-transparent clouds are present, LIS conservatively classifies them as cloud pixels, although they could be visually identified as snow [30].

C. Auxiliary data

We also use a digital elevation model (DEM) and a tree cover density map (TCD) in our processing workflow. The

DEM was sourced from the Copernicus global 30 m digital elevation model [31]. The TCD was obtained from the Copernicus Land Monitoring Service. It was derived from Sentinel-2 data and is available at 20 m resolution with pixel values ranging from 0% to 100% [32].

III. Methods

A. Sen-2-SPOT

To extract the snow cover area from SPOT images, we aim to use the U-Net architecture, a fully convolutional segmentation network which can be trained to extract contextual features from an image and localize them, generating an output image with pixel-wise annotation [25]. The U-Net should classify every valid pixel of a SPOT image in one of the following classes: ground (snow-free), snow or cloud. The main challenge is the lack of reference data to train the network since we do not have reference cloud/snow cover maps generated from SPOT images. We could find Landsat images (and their associated snow/cloud mask) acquired on the same day as some SPOT images, but even a time difference of a few minutes remains too large with respect to the variability of the cloud cover. Hence, to differentiate snow and cloud with the U-Net, it is imperative to generate reference data at the exact time the SPOT image was taken. As a solution, we developed a tool (Sen-2-SPOT) which emulates a SPOT data from a Sentinel-2 data, allowing us to make use of the corresponding Sentinel-2 cloud/snow cover map as reference data to train the U-Net from pseudo-SPOT data.

Sen-2-SPOT works in three main steps. First, it computes reflectance values for each pseudo-SPOT band by combining Sentinel-2 spectral band reflectances (Section III-A1). Then, it reduces the image dynamic by clamping the pseudo-SPOT reflectances between a minimum and maximum reflectance (Section III-A2). Last, it performs the radiometric compression to 8-bit encoding (Section III-A3). Sen-2-SPOT can emulate a SPOT image of any sensor from SPOT1 to SPOT5. It was designed to output image patches of any size, which can be directly ingested into the U-Net training.

Figure 1 schematizes the process of (i) emulating a pseudo-SPOT image patch from a Sentinel-2 image patch, (ii) passing the pseudo-SPOT patch through a U-Net to segment it into snow, ground and cloud classes using the Theia snow products as reference data to optimize the network weights.

1) Spectral bands combination: The spectral ranges of the HRV-IR bands overlap the spectral ranges of Sentinel-2. For example, the SPOT green band (0,54 - 0,58 μ m) overlaps both the Sentinel-2 blue band (0,46 - 0,525 μ m) and the Sentinel-2 green band (0,54 - 0,58 μ m) (see Fig. 2). Here, we use a first-order approach to estimate SPOT reflectance from Sentinel-2 data assuming a constant instrument sensitivity for each band. For each SPOT band, the pipeline finds the Sentinel-2 bands with overlapping spectral ranges and averages the corresponding reflectance using the overlapping band width as weight:

$$R_i = \frac{\sum_j X_{i,j} R_j}{\sum_j X_{i,j}} \tag{1}$$

Where i is the SPOT band to estimate and j an overlapping Sentinel-2 band. R is the reflectance value of a pixel and X_{ij} is the overlapping band width between i and j.

Only the Sentinel-2 red band overlaps the SPOT red band, so the reflectances are equal. The SPOT NIR band is overlapped by the Sentinel-2 bands 7 (vegetation), 8 (NIR) and 8A (narrow NIR) but, since the overlaps of 7 and 8A are covered by the overlap of 8, only the Sentinel-2 NIR band 8 is used. Hence, the value of the SPOT NIR band is equal to the one of the Sentinel-2 band 8.

2) Radiometric clamping: In contrast with images captured by modern sensors such as Sentinel-2 MSI, SPOT images are prone to saturation, especially over bright areas like clouds and snow covered surfaces. Given the lack of objective criteria to determine the saturation value, i.e. the reflectance above which a pseudo-SPOT image should be saturated, we adopted an empirical approach to estimate it from a large number of SPOT images. We parsed every SPOT product available over the Pyrenees from the SWH level 1C collection available and extracted the saturation values of the green, red and NIR reflectance bands from the metadata. The number of images is 271, 1281, 132, 776 and 683 for SPOT 1 to 5 respectively. A statistical model of the saturation values per spectral band was built separately for every SPOT sensor.

Figure 3 shows a scatter plot between the saturation values of the NIR and red bands for SPOT4. The data are also classified by bins of saturation values in the green band. The distribution follows multiple linear relationships, indicating that the saturation of the NIR and red bands are related, and the same observation can be drawn with the other band combinations. Therefore, we generate a trivariate Gaussian kernel density estimation (GKDE) from the distribution. We use the *gaussian_kde* function of the *scipy* Python library (Version 1.10.1) [33]. For a distribution of size n, the probability density function for a saturation vector of size $d = 3$ $x = (x_{green}, x_{red}, x_{nir})^T$ is written as:

$$\hat{f}(x) = \frac{1}{n|H|} \sum_{i=1}^{n} K(H^{-1}(x - x_i)) \quad \text{with} \quad H = hI_d \tag{2}$$

Where I_d is the identity matrix. *gaussian_kde* uses a Gaussian kernel K to smooth the data, and a bandwidth parameter h to determine the width of the kernel. The bandwidth controls the trade-off between bias and variance in the estimation, with a smaller bandwidth leading to a higher variance and a larger bandwidth resulting in a higher bias.

Figure 3 (center) displays the 2D projection of the density function between NIR and red values. We use a small bandwidth = 0.1 to ensure that the GKDE will fit the shapes of the multiple linear relations and keep a probability of 0 between them. When processing a Sentinel-2 image, the Sen-2-SPOT pipeline will randomly sample a set of saturation values from this probability density. Fig. 3 (right) illustrates how the sampling of 500 data points reproduces a similar distribution to that of SPOT4.

We apply the same process to estimate the minimum reflectance values from the metadata files and generate a

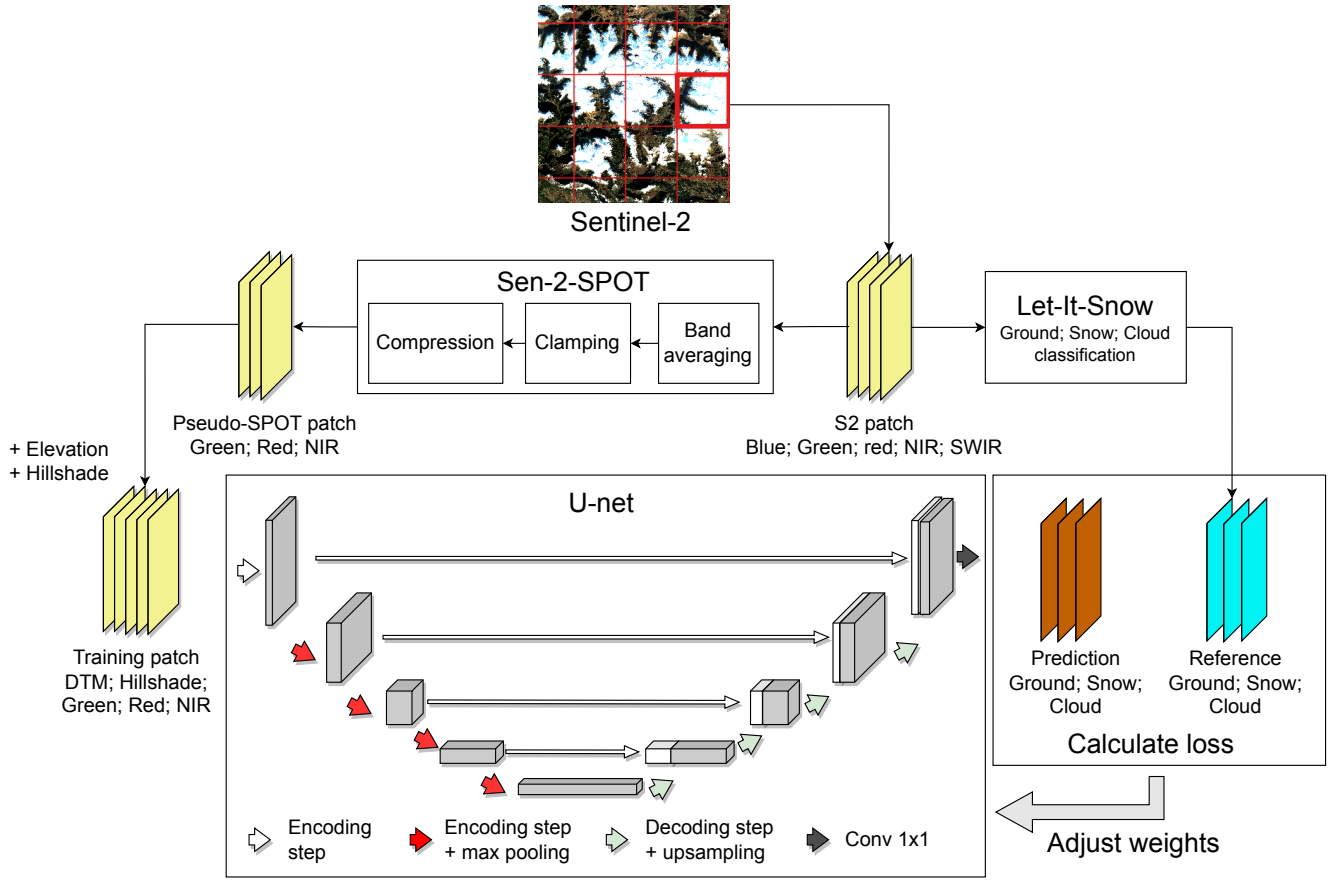

Fig. 1. Flow chart of Sen2SPOT + U-Net training.

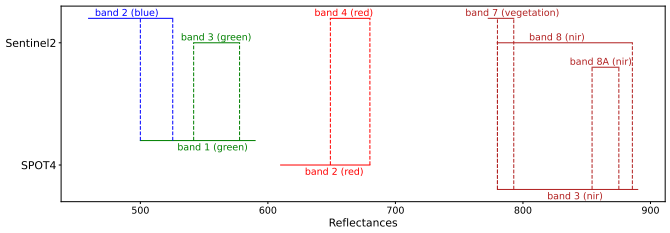

Fig. 2. SPOT4 spectral bands and corresponding Sentinel-2 spectral bands.

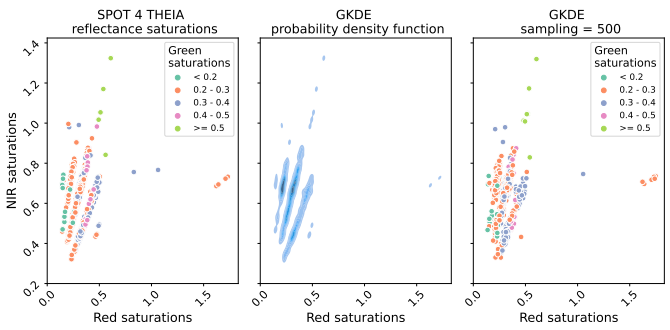

Fig. 3. Distribution of SPOT4 saturation reflectance values (green, red, nir). Left: distribution of values from an actual sample of images over the Pyrenees; center: gaussian kernel density estimation of the distribution; right: 500 random samplings from the estimation.

synthetic minimum value for each band and sensor. However, the minimum reflectance exhibits much less variability than the maximum reflectance value and its estimation is far less critical for the snow/cloud classification.

3) Compression: For each Sentinel-2 band, Sen-2-SPOT re-scales the clamped reflectance values R_{S2} into a distribution R_{SPOT} of 256 different reflectance values inside the range $[R_{min}, R_{max}]$ (Eq. 3) with R_{min} and R_{max} the minimum and maximum reflectance values of R_{S2} III-A2.

$$R_{8bit} = \lfloor 255 \frac{R_{S2} - R_{min}}{R_{max} - R_{min}} \rceil \quad (3)$$

Where the $\lfloor a \rceil$ notation denotes the function which rounds a decimal number a to the nearest integer.

Then, the 8 bits values R_{8bit} are rescaled to integers in milli-reflectance units to match the distribution format of SWH level 1C images (Section II).

$$R_{SPOT} = 1000 \frac{R_{max} - R_{min}}{255} R_{8bit} + R_{min} \quad (4)$$

Figure 4 shows the conversion from an (a) input Sentinel-2 patch with a histogram (b) of the red band to (c) a pseudo-SPOT4 patch with a histogram (d) of the same band. In this example, red reflectance values are converted from a range of [0.05; 1.30] to [0.05; 0.41], which creates large areas of saturation.

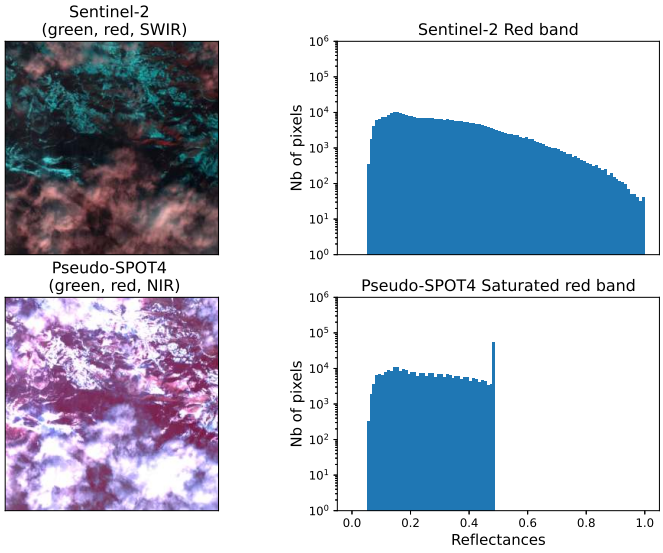

Fig. 4. Above: Sentinel-2 image patch with a histogram of the red band. Below: saturated pseudo-SPOT4 image patch with a histogram of the red band.

B. U-Net

1) Architecture: We use the same U-Net architecture as described in [25]. A U-Net is composed of two symmetric paths for encoding and decoding. At the encoding path, the network takes an input patch with C channels and passes it through five steps of convolution with a downsampling using a 2×2 maximum pooling between each step. An encoding step is described as:

- A first convolution layer with $64 \times 2^i \times C_i$ kernels of size 3x3 bringing the number of channels to $C_{i+1} = 64 \times 2^i$. Where $i \in [0..4]$ is the step number.
- A batch normalization
- A rectified linear unit (ReLU) activation function, bringing negative values to zero
- A repeat of the three previous steps

The decoding path consists of four steps of convolution with a bilinear upsampling followed by a skip connection with the encoder before each step. A skip connection involves concatenating the output of the convolution layers from an encoding step with the features generated by the corresponding decoding input, enabling the transfer of raw information from the former to the latter (as illustrated by the crop and copy operations in Fig. 1). A decoding step brings the number of channels to $C_{i+1} = 1024/2^i$ (with $i \in [0..3]$ the decoding step number). The last step ends with a final convolution layer with a 1×1 kernel generating a patch with one channel for each class.

During the training phase, it is important to ensure that input data is normalized or standardized to a consistent scale before being passed through the network. Hence, for each channel, we normalize the dataset used for the training by subtracting the mean from band and dividing the result by the standard deviation. In addition, because dense tree canopy impedes snow detection performances with optical satellite sensors [34], we mask pixels with a tree cover density greater than 50%.

2) Input features: An input patch is formed by the concatenation of the pseudo-SPOT image green, red and NIR bands, and two additional bands: a digital elevation model (DEM) and a hillshade layer giving the pixel illumination at the acquisition time. The DEM was included because the snow cover area is largely influenced by elevation. The hillshade was included because the surface reflectance of the snow cover at the acquisition time is highly correlated to the hillshade contrary to clouds. By providing these contextual features, the model should be better equipped to discriminate cloud and snow areas.

The elevation is extracted from the Copernicus DEM and resampled to 20 m using bilinear interpolation. The hillshade was calculated from the DEM and from the sun azimuth (ϕ) and zenith (θ) angles at acquisition time given in the image metadata:

$$Hillshade_{pixel} = 255(\cos\theta \cos S_{pixel} + \sin\theta \sin S_{pixel} \cos(\phi - A_{pixel})) \quad (5)$$

$$S_{pixel} = \arctan\sqrt{(\frac{dz}{dx})^2 + (\frac{dz}{dy})^2} \quad (6)$$

$$A_{pixel} = \arctan2(\frac{dz}{dy} - \frac{dz}{dx}) \quad (7)$$

The slope S of a pixel is calculated from the horizontal $\frac{dz}{dx}$ and vertical $\frac{dz}{dy}$ rates of elevation change in a 3-by-3 window centered on the pixel. The aspect A is the orientation of the slope.

3) Loss: Table II shows how the classes can be unbalanced in the training dataset, with the snow class being the minority. Hence, when comparing the output of the U-Net with the reference data, we calculate the loss for one batch using a weighted cross entropy function :

$$L = \sum_{i=1}^{I} \sum_{j=0}^{2} W_j R_{i,j} \log(P_{i,j}) \quad (8)$$

$$\text{with} \quad W_j = \frac{\#train}{\#train_j} \quad (9)$$

With $j \in [0, 1, 2]$ the class (3 being a mask), i the pixel, W the weight, R the target (true) probability 1 or 0, P the predicted probability, and I the amount of pixels in the batch. We calculate the weight for the class j by dividing the number of pixels in the training dataset with the number of pixel of class j in the training dataset.

4) Training: An input patch has a size $N \times N \times C$ with N = 572 pixels and C = 5 channels (3 pseudo-SPOT bands, 1 DTM band and 1 hillshade band). The U-Net is trained iteratively with batches of 11 patches at each iteration. The training dataset is run through the U-Net multiple times (one time being an epoch) in random order.

For each model, 90% of the training dataset is used to actually train the U-Net while the other 10% is used to follow the evolution of the loss after each epoch. The initial

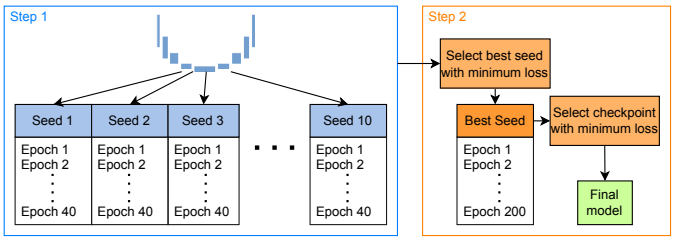

Fig. 5. Step 1: light training of the U-Net model. Step 2: heavy training from the best checkpoint of step 1

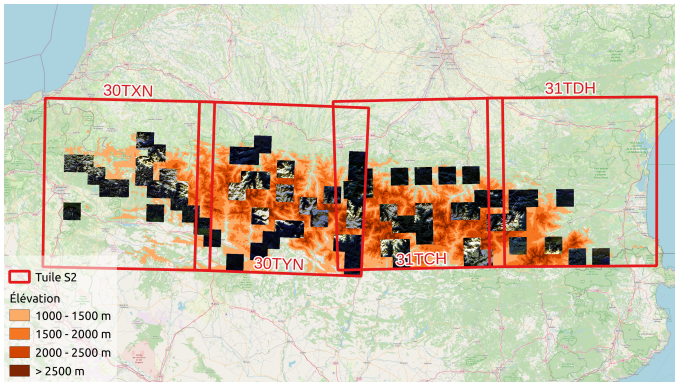

Fig. 6. Distribution of 80 Sentinel-2 patches across the Pyrenees (> 1000 m)

parameters of a neural network impact if and how well the training will converge toward a minimum in the loss curve. In this work, the training is done in two steps (Fig. 5).

- The first step is a preliminary short training repeated 10 times with the same split of the training set but with different weights initialization (seeds). The trainings last 40 epochs each and use a learning rate of 0.01. At the end of each epoch we save a checkpoint of the network's weights and biases and of the training and evaluation losses. In this step, we use a high learning rate to "look" for the checkpoint where the U-Net can converge to a state with minimum loss.
- The second step is a more intensive and time-consuming training starting from the checkpoint found in step 1. Step 2 uses a lower learning rate of 0.0001 for 200 epochs. A checkpoint is saved after each epoch and we keep the one with the lowest evaluation loss.

C. Evaluation of the model

1) Evaluation of the model training strategy: Using Sen-2-SPOT allows us to test several strategies to train the model. In particular, we intend to evaluate the best strategy to obtain a robust model to classify SPOT images over a large region such as the Pyrenees mountain range.

We evaluate five models: four "local" models which are trained and tested separately from data extracted over one of the Sentinel-2 tiles covering the French Pyrenees (30TXN, 30TYN, 31TCH and 31TDH) and another model (referred to as PYR) trained and tested from the combination of the same four training datasets and therefore covering the whole French Pyrenees (PYR) (Table II).

To train and test these models, we randomly select 12 Sentinel-2 images per tile, i.e. one image per month between 2016 and 2019, for a total of 46 images. The random selection was done with the following conditions: (i) a minimum of ten days between two images of the same tile, (ii) no simultaneous acquisitions between two adjacent tiles to avoid duplicate data in the overlapping regions, (iii) a cloud coverage below 90%, and (iv) the relative orbit numbers of the Sentinel-2 images for the tiles 30TXN, 30TYN, 31TCH and 31TDH are respectively 94, 51, 51 and 8. That last condition ensures that only complete images (no nodata in the image edges) are selected.

We divided the input Sentinel-2 data into patches. However, the output patches from the U-Net will have a reduced size of 388 pixels due to repeated convolutions. Consequently, patches are extracted from input Sentinel-2 images with an overlap of 184 pixels between them.

To reduce the amount of unnecessary data outside of the Pyrenees (without any mountainous features relevant for the training) we use the DEM resampled to 20 m to filter the patches and only keep those covering or partially covering regions above 1200 meters of altitude. As an example, Fig. 6 shows 80 patches distributed among four Sentinel-2 products (one for each tile). Tiles 30TXN and 31TDH, situated at the extremities, cover both smaller portions of the Pyrenees above 1200 meters. Table II shows the distribution of unmasked snow and cloud pixels used in the training and testing sets between the four tiles. Even after filtering for high elevations, the total distribution is still uneven with approximately half the pixels being either snow or clouds and two times more clouds than snow pixels. Tiles 30TXN and especially 31TDH have a noticeably smaller sample size (Fig. 6).

We randomly split the data into two sets, which are used to train and test the U-Net model with pseudo-SPOT data, with the respective ratios of 80% and 20%. These ratios are imposed for each Sentinel-2 image. We evaluate each model by computing the confusion matrix metrics for the snow, cloud and ground labels using those 20% testing data. Given that the testing datasets are greatly imbalanced, the F1 score is a suitable measure in such cases for evaluating a classifier's performances for a specific label. For each tile, we computed the F1 scores of each label snow, cloud and ground. We also computed the precison and recall scores of the snow label for each month of the year. The Cohen's kappa coefficient expresses with only one value the overall agreement between two classifications. While the kappa might not be as useful as other methods when comparing different datasets [35] or when evaluating a classifier on an imbalanced dataset, we use it to compare, for each tile, the local model and the PYR model.

2) Evaluation using actual SPOT data: In the previous section we rely on Sentinel-2 data for training and validation. To make sure that the Sen-2-SPOT/U-Net pipeline is efficient to classify actual SPOT data, we create a validation dataset from actual SPOT4 data. Using SPOT4 allows us to generate a reliable reference dataset by supervised classification of the image since SPOT4 images have a SWIR band. For the same Sentinel-2 tiles as above, we select four level 1C SPOT4

TABLE II
DISTRIBUTION OF TRAINING (TOP TABLE) AND TESTING (BOTTOM TABLE) PIXELS ACROSS THE FOUR TILES 30TXN, 30TYN, 31TCH, 31TDH OVER THE PYRENEES

Training datasets	Tiles				Total
	30TXN	30TYN	31TCH	31TDH	
Nb of pixels	5.2e7	10.9e7	8.5e7	3e7	27.7e7
% of snow pixels	11.8	16	23.8	14.6	17.4
% of cloud pixels	28.6	35.5	29.1	26	31.2
Testing dataset					
Nb of pixels	1.3e7	2.9e7	2.2e7	0.8e7	7.2e7
% of snow pixel	14.3	16.6	23.9	12.1	17.9
% of cloud pixel	28.2	34.1	31.8	31.6	32

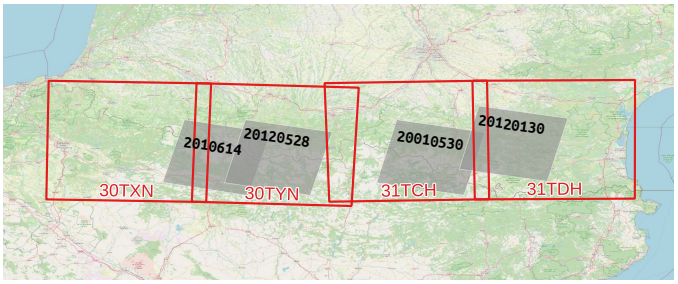

Fig. 7. Map of the four SPOT4 images used for the inference test.

products (Fig. 7). The selected products contain a mixture of snow, ground and cloud pixels, including regions where clouds and snow are overlapping. Figure 8 shows an example of how the SWIR band (b) helps differentiate between clouds (white color) and snow (cyan color), as both the snow and cloud covers are entirely saturated in the green and red bands as seen from a manually generated saturation map (a). In (c), the SWIR band is replaced with the NIR band and both clouds and snow are the same color.

The supervised classification of the selected SPOT4 images is done by manual selection of snow, cloud and ground samples from color composites using the SWIR band. We use 50% of the samples to run a Gaussian mixture model to classify the entire image. This procedure is repeated for each SPOT4 product. The classification is evaluated both visually and using the confusion matrix computed from the remaining 50% samples that were not used for the training (results not shown here for the sake of brevity). If necessary, the classification is refined by adding new samples. Table III shows the distribution of classified snow and cloud pixels for each tile.

Heavy saturation, transparent clouds and SPOT4 images being encoded in 8 bit pixels made the classification challenging. A visual inspection of the four images has shown that this is especially the case in areas of transition between snow and ground and between snow and clouds where, in both cases, pixels tends to be miss-classified as clouds. On the other hand, pixels are rarely miss-classified as snow. For this reason, we focus this evaluation of the U-Net's performances on the classification of snow pixels only. For each image we compute the Precision and Recall of the snow label.

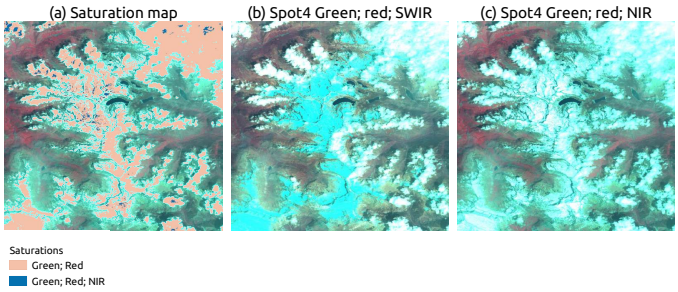

Fig. 8. Zoom on the 20120528 SPOT4 image with (a) a manually generated saturation map (b) a false color with the Green, Red and SWIR bands (c) the Green, Red, and NIR bands.

TABLE III
DISTRIBUTION OF SPOT4 PIXELS ACROSS THE FOUR TILES 30TXN, 30TYN, 31TCH, 31TDH OVER THE PYRENEES

SPOT4 dataset	Tiles				Total
	30TXN	30TYN	31TCH	31TDH	
Nb of pixels	1.1e7	3.7e7	2.8e7	1.8e7	9.3e7
% of snow pixel	6.5	13.0	21.1	73.6	25.8
% of cloud pixel	3.4	33.2	10.7	4.4	18.1

IV. RESULTS

A. Evaluation of the model training strategy

Figure 9 shows, for the local and PYR models respectively, the confusion matrix between the reference labels from LIS and the predicted labels from the U-Net on the pseudo-SPOT datasets over the combined four Pyrenean tiles.

Both (a) and (c) show similar performances: for each label, the amounts of true snow, true ground and true clouds are one or two orders of magnitude larger than the amount of false detections. In (a), between snow pixels and cloud pixels, the U-Net has twice the amount of falsely predicted snow than of falsely predicted cloud.

Figures 9 (b) and (d) show, for the local and PYR models respectively, the kappa and F1 scores for each tile. Despite a larger training dataset, the PYR model only brings a slight increase of both kappa and F1 performances for the 30TXN, 30TYN, and 31TCH tiles. On the contrary, we notice a more marked improvement over the 31TDH tile, which has the lowest performances with both the kappa, the snow F1 and the cloud F1 scores below 0.75. As seen in Table II, the 31TDH training dataset is notably smaller than the other datasets, which reduces the effectiveness of the locally trained model 31TDH while the PYR model relies more on its training over the other tiles.

The tile 30TYN also shows significantly lower scores than 30TXN and 31TCH despite being the tile with the largest dataset. To better understand the factors behind the results of Fig. 9, Fig. 10 shows, for each month of the year and for each tile, the kappa, snow precision score and snow recall score of the PYR model. The snow scores have less weight on the kappa in the warmer months (June to October) as there is little snow compared to the rest of the year (a negligible amount of snow can produce extremely low or high scores).

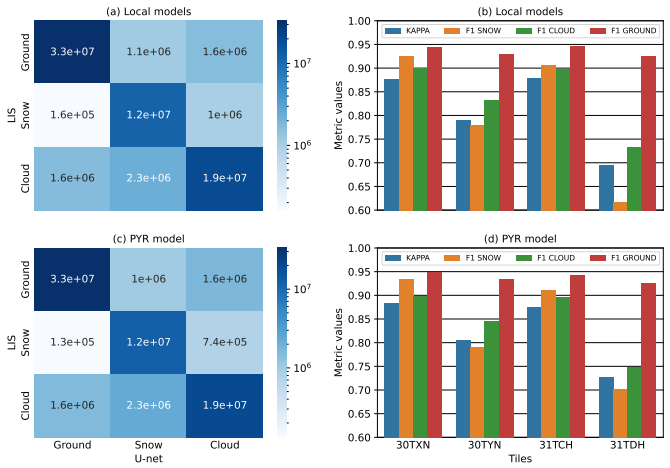

Fig. 9. For the ground, snow and cloud labels, confusion matrices (left) and per-tile performance metrics (right) of the local models 30TXN, 30TYN, 31TCH, and 31TDH (a)(b) and of the PYR model (c)(d) on the testing pseudo-SPOT4 dataset.

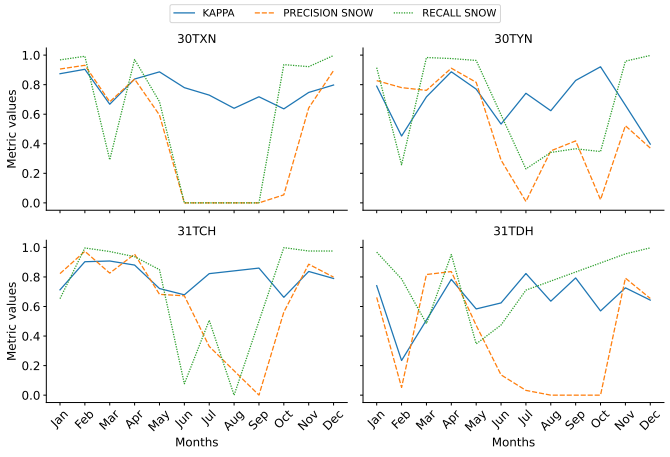

Fig. 10. Precision and recall for the snow class and kappa score of the PYR model over each tile for each month of the year.

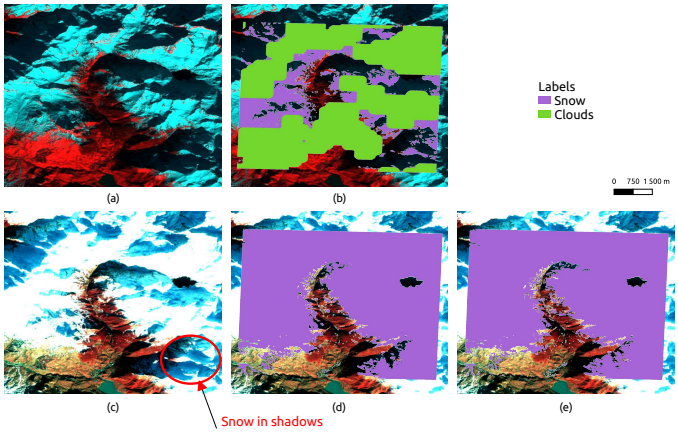

Fig. 11. Patch of December over the 30TYN tile: (a) False color green, red, SWIR of the Sentinel-2 patch (b) LIS classification (c) False saturated color green, red, NIR of the pseudo-SPOT4 patch (d) U-Net classification with the local model (e) U-Net classification with the PYR model.

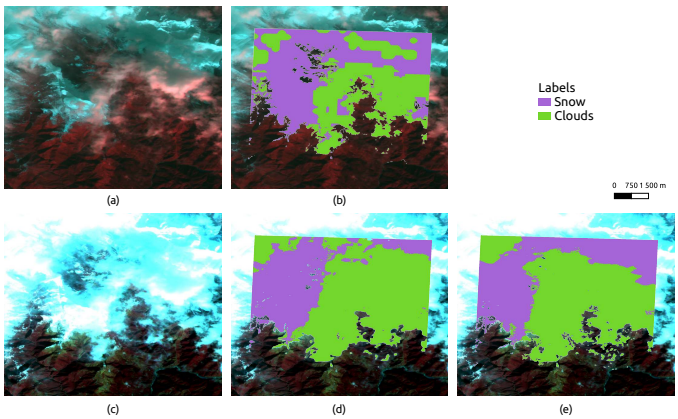

Fig. 12. Patch of February over the 30TYN tile: (a) False color green, red, SWIR of the Sentinel-2 patch (b) LIS classification (c) False saturated color green, red, NIR of the pseudo-SPOT4 patch (d) U-Net classification with the local model (e) U-Net classification with the PYR model.

In colder months, low precision scores happen when the U-Net retrieves snow pixels in regions where LIS did not detect them. This happens especially when LIS detected clouds where there are none, for example in February for the tile 31TDH and in December for the tile 30TYN. Figure 11 shows a patch of the December product of tile 30TYN, where both the local and PYR models retrieve snow pixels (purple) from falsely detected clouds (green) and from missed snow cover in shadow areas (red circle).

Where clouds are present, the U-Net generates larger cloud covers in comparison to LIS, reducing the recall score of the snow label. Figure 12 shows a patch of the February product of tile 30TYN, where both the local and PYR models have a coarser cloud cover than LIS.

B. Evaluation using actual SPOT data

Figures 13 (a) and (c) show, for the local and PYR models respectively, the confusion matrix between the reference labels generated with manual classification and the predicted labels from the U-Net on the SPOT4 datasets over the combined four Pyrenean tiles. For (a), we ran the local models over their respective tile and summed the respective confusion matrices together. The recall and precision of the snow label are respectively 0.46 and 0.97 for the local models and respectively 0.44 and 0.98 for the PYR model.

From (b) and (d), where each tile correspond to one SPOT4 image, we do not observe the same performance improvements from the PYR model over SPOT4 images as we did over pseudo-SPOT4 images. With the exception of 30TYN, the PYR model improves the snow precision slightly across the dataset at the cost of a loss of snow recall. Figures 14, 15, and 16 show comparative maps of the manual classification, local model prediction and PYR model prediction for, respectively, the tiles 30TXN, 30TYN, and 31TCH. We also display the TCD masks, as the pixels corresponding to a TCD $>$ 50% were not included in the analysis. In each case, the U-Net enlarges the detected clouds, including the cloud shadows, with a buffer. There is a stronger effect from the PYR model, losing both true and false snow pixels which explains the

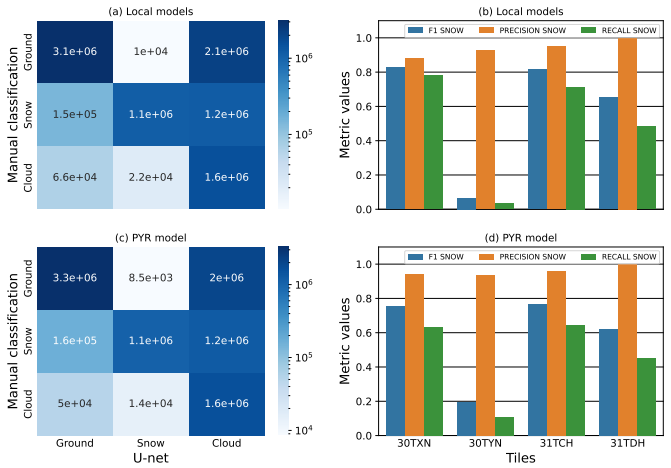

Fig. 13. Ground, snow and clouds labels confusion matrices (left) and per-tile snow label performance metrics (right) of the local models 30TXN, 30TYN, 31TCH, and 31TDH (a)(b) and of the PYR model (c)(d) on the testing SPOT4 dataset.

improvement of the snow precision and loss of snow recall. The tile 31TDH is not displayed as it is heavily masked by high TCD and does not provide additional information to the results:

- **30TXN**: Figure 14 displays an image with sparse clouds above and next to snow-covered areas. The local model correctly classifies most of the snow with a precision of 0.88 and recall of 0.78. The PYR model has a snow precision of 0.94 and a snow recall of 0.63.
- **30TYN**: Figure 15 displays a heavily saturated image with a mix of clouds and snow difficult to differentiate without the SWIR band. Compared to the reference, the U-Net classifies most of the image as clouds, resulting in a very low recall of 0.034 (local). The PYR model retrieves more snow without error, augmenting the recall to 0.1 without lowering the precision (both models at 0.93).
- **31TCH**: We observe a behavior similar to 30TXN. Figure 16 displays an image with clouds mostly above borders between snow and no-snow areas. The local model correctly classifies most of the snow with a precision of 0.95 and recall of 0.71. The PYR model has a snow precision of 0.96 and a snow recall of 0.64.

V. Discussion

When tested on pseudo-SPOT4 images, both the local and PYR models showed high agreement with LIS, with the lowest kappa being close to 0.7. Those results show that the less computationally intensive local models are sufficient for tiles with significant snow cover. The PYR model trained over the entire French Pyrenees only improves the results on tile 31TDH where snow pixels are rare. Yet, the performance on this tile remains lower than the performance of the other tiles, suggesting that the information brought from the other tiles does not completely compensate for the lack of local training data.

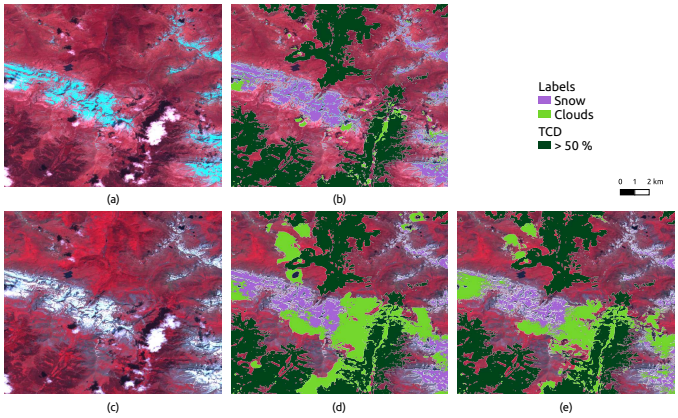

Fig. 14. SPOT4 20100614 image section over the 30TXN tile: (a) False color green, red, SWIR (b) Supervised classification (c) False color green, red, NIR (d) U-Net classification with the local model (e) U-Net classification with the PYR model.

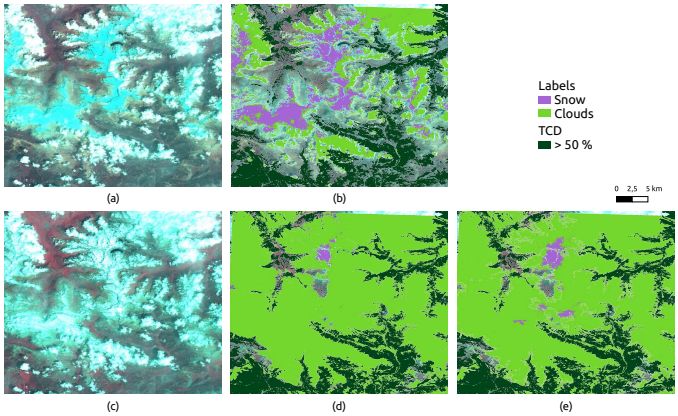

Fig. 15. SPOT4 20120528 image section over the 30TYN tile: (a) False color green, red, SWIR (b) Supervised classification (c) False color green, red, NIR (d) U-Net classification with the local model (e) U-Net classification with the PYR model.

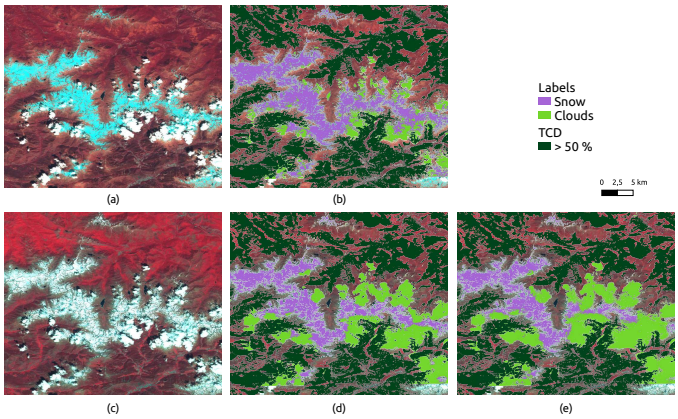

Fig. 16. SPOT4 20010530 image section over the 31TCH tile: (a) False color green, red, SWIR (b) Supervised classification (c) False color green, red, NIR (d) U-Net classification with the local model (e) U-Net classification with the PYR model.

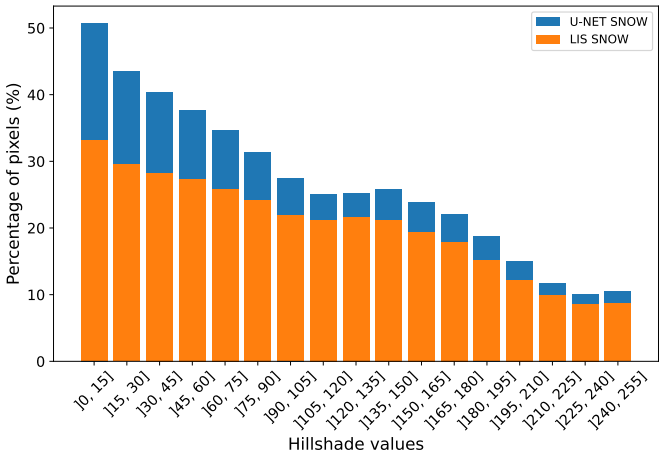

Fig. 17. For the 30TYN pseudo-SPOT4 testing dataset, bar plots of the local model (blue) and LIS (orange) of the percentage of pixels detected as snow across the hillshade values.

The LIS algorithm (as discussed in section II-B), which generates snow and cloud maps at the operational scale from Sentinel-2 data, was designed to minimize false snow positives at the cost of overestimating the cloud mask since cloud pixels can be filled by temporal interpolation [18], [30]. The U-Net reproduces this behavior, with an overall precision of 95% at the cost of overestimating the amount of clouds. When tested on independent and actual SPOT4 data, we also observe that the local models yield high precisions in snow detection while overestimating the cloud cover. However, the PYR model does not improve the performances over the SPOT4 dataset as it did over the pseudo-SPOT4 dataset, suggesting that a locally trained model might be a good approach to process more SPOT images.

Another interesting observation from the results is the detection of snow pixels by the U-Net in less illuminated areas compared to LIS. From the pseudo-SPOT4 testing 30TYN dataset, Fig. 17 shows for both the local U-Net model (blue) and LIS (orange) a bar plot of the percentage of pixels detected as snow across 17 equal ranges of hillshade values. The local model detects more snow pixels than LIS, and the difference gets bigger as the hillshade value gets lower. Inside the]0,15] range, hillshade values of 1 corresponds to areas with no direct sunlight. In those areas the U-Net model detected 1.5 times more snow pixels than LIS. Those results hint at the potential capacity of the model to retrieve missing snow pixels in mountain shadows (as seen in Fig. 11) as an improvement over the reference data.

VI. Conclusion

Emulating SPOT images from Sentinel-2 data (Sen-2-SPOT) enables to train a deep learning model (U-Net) to extract the snow cover area from historical SPOT images. We find that the trained U-Net performs well despite the frequent saturation of SPOT bands over cloud and snow surface, and the absence of a SWIR band (SPOT1-3). This is due to the fact that the U-Net takes advantage of the spatial context around each pixel to determine its class in contrast with standard pixel-based classifications. The training can be performed separately by Sentinel-2 tile, although a more robust model might be obtained in areas with scarce snow cover by using neighbouring tiles. The U-Net outputs a confidence value for the detected class, and that confidence value can potentially be used to generate a pixel quality flag with the snow map. However, more work should be done on the interpretation of those values, as high confidence values are often given to false clouds.

We performed an additional test (not shown here) by removing the green and red (which are the most prone to saturation) and keeping only the NIR band, elevation and hillshade with the same dataset. The results show that this reduces the number of false cloud pixels (and therefore increases true snow). But it also increases the number of false ground in snow covered areas, which is more problematic for future analysis of the snow cover maps.

In theory, this approach has no geographical limitations since Sentinel-2 can be used to produce snow cover maps anywhere. However, the method does not work in forest regions due to the inability of optical sensors to observe the snow underneath dense canopies. In addition, it was designed and tested for mountain regions where the elevation plays a first-order role in the spatial distribution of the snow cover at the pixel level. It should be less effective in flat regions where the pixel elevation has less importance than the subpixel topography like arctic regions. It may be possible to further improve the model by feeding additional auxiliary data to the U-Net such as snow cover probability maps or gridded meteorological data.

Once the training is done, the inference is fast and can be applied to large volumes of data. Therefore, it can be used to generate high resolution time series of snow cover maps from the SPOT World Heritage collection over entire mountain ranges like the Pyrenees or the Alps. This pipeline can also be applied to reconstruct the snow cover from other satellite missions with similar spatial and spectral characteristics like Landsat.

Acknowledgment

This work was supported by the ANR TOP project, grant ANR-20-CE32-0002 of the French Agence Nationale de la Recherche. We thank Olivier Hagolle and Jean-Pierre Dedieu for their insightful comments and suggestions on this work. We also thank Philippe Choler for the coordination of the ANR project.

Appendix A

References

[1] C. Fierz, A. R.L, D. Y, E. P, E. Greene, D. Mcclung, K. Nishimura, P. Satyawali, and S. Sokratov, *The international classification for seasonal snow on the ground (UNESCO, IHP (International Hydrological Programme)–VII, Technical Documents in Hydrology, No 83; IACS (International Association of Cryospheric Sciences) contribution No 1)*, Jan. 2009.

[2] A. C. Edwards, R. Scalenghe, and M. Freppaz, "Changes in the seasonal snow cover of alpine regions and its effect on soil processes: a review," *Quaternary international*, vol. 162, pp. 172–181, 2007.

[3] T. Jonas, C. Rixen, M. Sturm, and V. Stoeckli, "How alpine plant growth is linked to snow cover and climate variability," *Journal of Geophysical Research: Biogeosciences*, vol. 113, no. G3, 2008.

[4] P. Choler, "Growth response of temperate mountain grasslands to inter-annual variations of snow cover duration," *Biogeosciences Discussions*, vol. 12, pp. 3025–3055, Jun. 2015.

[5] J. Revuelto, D. Gómez, E. Alonso-González, I. Vidaller, F. Rojas-Heredia, C. Deschamps-Berger, J. García-Jiménez, G. Rodríguez-López, J. Sobrino, R. Montorio *et al.*, "Intermediate snowpack melt-out dates guarantee the highest seasonal grasslands greening in the pyrenees," *Scientific Reports*, vol. 12, no. 1, p. 18328, 2022.

[6] L. Francon, E. Roussel, J. Lopez-Saez, M. Saulnier, M. Stoffel, and C. Corona, "Alpine shrubs have benefited more than trees from 20th century warming at a treeline ecotone site in the French Pyrenees," *Agricultural and Forest Meteorology*, vol. 329, p. 109284, Feb. 2023.

[7] G. Blöschl, "Scaling issues in snow hydrology," *Hydrological Processes*, vol. 13, no. 14-15, pp. 2149–2175, 1999.

[8] T. R. Loveland and J. R. Irons, "Landsat 8: The plans, the reality, and the legacy," *Remote Sensing of Environment*, vol. 185, pp. 1–6, Nov. 2016.

[9] J. Ju and D. P. Roy, "The availability of cloud-free Landsat ETM+ data over the conterminous United States and globally," *Remote Sensing of Environment*, vol. 112, no. 3, pp. 1196–1211, Mar. 2008.

[10] Y. Zhang, C. E. Woodcock, P. Arévalo, P. Olofsson, X. Tang, R. Stanimirova, E. Bullock, K. R. Tarrio, Z. Zhu, and M. A. Friedl, "A Global Analysis of the Spatial and Temporal Variability of Usable Landsat Observations at the Pixel Scale," *Frontiers in Remote Sensing*, vol. 3, 2022.

[11] J. Parajka and G. Blöschl, "Spatio-temporal combination of MODIS images – potential for snow cover mapping," *Water Resources Research*, vol. 44, no. 3, 2008.

[12] S. Gascoin, O. Hagolle, M. Huc, L. Jarlan, J.-F. Dejoux, C. Szczypta, R. Marti, and R. Sánchez, "A snow cover climatology for the Pyrenees from MODIS snow products," *Hydrol. Earth Syst. Sci.*, vol. 19, no. 5, pp. 2337–2351, May 2015.

[13] J.-P. Dedieu, B. Z. Carlson, S. Bigot, P. Sirguey, V. Vionnet, and P. Choler, "On the Importance of High-Resolution Time Series of Optical Imagery for Quantifying the Effects of Snow Cover Duration on Alpine Plant Habitat," *Remote Sensing*, vol. 8, no. 6, p. 481, Jun. 2016.

[14] M. Abrams, S. Hook, and B. Ramachandran, "Aster user handbook, version 2," *Jet propulsion laboratory*, vol. 4800, p. 135, 2002.

[15] J. Nosavan, A. Moreau, A. Masse, B. Chausserie-Laprée, and C. Caillet, *Spot world heritage: spot 1-5 data curation and valorization with new enhanced swh products*, Nov. 2017.

[16] Z. Barrou Dumont, S. Gascoin, and J. Inglada, "Contribution de SPOT World Heritage aux séries temporelles d'observation satellitaires des montagnes françaises," *Revue Française de Photogrammétrie et de Télédétection*, vol. 225, no. 1, pp. 1–8, Feb. 2023.

[17] Z. Zhu, S. Wang, and C. E. Woodcock, "Improvement and expansion of the Fmask algorithm: cloud, cloud shadow, and snow detection for Landsats 4–7, 8, and Sentinel 2 images," *Remote sensing of environment*, vol. 159, pp. 269–277, 2015.

[18] S. Gascoin, M. Grizonnet, M. Bouchet, G. Salgues, and O. Hagolle, "Theia Snow collection: high-resolution operational snow cover maps from Sentinel-2 and Landsat-8 data," *Earth System Science Data*, vol. 11, no. 2, pp. 493–514, Apr. 2019.

[19] J. Dozier, "Spectral signature of alpine snow cover from the landsat thematic mapper," *Remote Sensing of Environment*, vol. 28, pp. 9–22, Apr. 1989.

[20] T. H. Painter, J. Dozier, D. A. Roberts, R. E. Davis, and R. O. Green, "Retrieval of subpixel snow-covered area and grain size from imaging spectrometer data," *Remote Sensing of Environment*, vol. 85, no. 1, pp. 64–77, Apr. 2003.

[21] T. Stillinger, D. A. Roberts, N. M. Collar, and J. Dozier, "Cloud Masking for Landsat 8 and MODIS Terra Over Snow-Covered Terrain: Error Analysis and Spectral Similarity Between Snow and Cloud," *Water Resources Research*, vol. 55, no. 7, pp. 6169–6184, 2019.

[22] Y. Zhan, J. Wang, J. Shi, G. Cheng, L. Yao, and W. Sun, "Distinguishing Cloud and Snow in Satellite Images via Deep Convolutional Network," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 10, pp. 1785–1789, Oct. 2017, conference Name: IEEE Geoscience and Remote Sensing Letters.

[23] Y. Lu, T. James, C. Schillaci, and A. Lipani, "Snow detection in alpine regions with Convolutional Neural Networks: discriminating snow from cold clouds and water body," *GIScience & Remote Sensing*, vol. 59, no. 1, pp. 1321–1343, Dec. 2022.

[24] X. X. Zhu, D. Tuia, L. Mou, G.-S. Xia, L. Zhang, F. Xu, and F. Fraundorfer, "Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 8–36, Dec. 2017.

[25] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, ser. Lecture Notes in Computer Science, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. Cham: Springer International Publishing, 2015, pp. 234–241.

[26] J. H. Jeppesen, R. H. Jacobsen, F. Inceoglu, and T. S. Toftegaard, "A cloud detection algorithm for satellite imagery based on deep learning," *Remote Sensing of Environment*, vol. 229, pp. 247–259, Aug. 2019.

[27] M. Drusch, U. Del Bello, S. Carlier, O. Colin, V. Fernandez, F. Gascon, B. Hoersch, C. Isola, P. Laberinti, P. Martimort, A. Meygret, F. Spoto, O. Sy, F. Marchese, and P. Bargellini, "Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services," *Remote Sensing of Environment*, vol. 120, pp. 25–36, May 2012.

[28] O. Hagolle, M. Huc, D. V. Pascual, and G. Dedieu, "A multi-temporal method for cloud detection, applied to FORMOSAT-2, VENμS, LANDSAT and SENTINEL-2 images," *Remote Sensing of Environment*, vol. 114, no. 8, pp. 1747–1755, Aug. 2010.

[29] Z. Barrou Dumont, S. Gascoin, O. Hagolle, M. Ablain, R. Jugier, G. Salgues, F. Marti, A. Dupuis, M. Dumont, and S. Morin, "Brief communication: Evaluation of the snow cover detection in the Copernicus High Resolution Snow & Ice Monitoring Service," *The Cryosphere*, vol. 15, no. 10, pp. 4975–4980, Oct. 2021.

[30] Z. Barrou Dumont and S. Gascoin, "Evaluation of a High-Resolution Operational Snow Cover Area Classification Algorithm," in *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, Jul. 2021, pp. 5485–5488.

[31] European Space Agency and Airbus, "Copernicus DEM," 2022, institution: European Space Agency. [Online]. Available: https://spaceata.copernicus.eu/collections/copernicus-digital-elevation-model

[32] C. Sannier, J. Gallego, T. Langanke, U. Donezar, and A. Pennec, "Tree cover area estimation in europe based on the combination of in situ reference data and the copernicus high resolution layer on tree cover density," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XLVIII-M-1-2023, pp. 277–284, 2023.

[33] D. W. Scott, *Multivariate Density Estimation: Theory, Practice, and Visualization*. John Wiley & Sons, Mar. 2015, google-Books-ID: pIAZBwAAQBAJ.

[34] A. Muhuri, S. Gascoin, L. Menzel, T. S. Kostadinov, A. A. Harpold, A. Sanmiguel-Vallelado, and J. I. López-Moreno, "Performance assessment of optical satellite-based operational snow cover monitoring algorithms in forested landscapes," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 7159–7178, 2021.

[35] R. G. Pontius and M. Millones, "Death to Kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment," *International Journal of Remote Sensing*, vol. 32, no. 15, pp. 4407–4429, Aug. 2011.