CloudSim: a fair benchmark for comparison of methods for times series reconstruction from cloud and atmospheric contamination

Julien, Y., Sobrino, J. A.
Global Change Unit,
Image Processing Laboratory,
University of Valencia, Spain.
yves.julien@uv.es

Abstract— **Cloud contamination of optical data is a constant and annoying feature of time series analyses, whether while using vegetation indices or surface temperatures, since it tends to decrease artificially the values taken by these parameters. Therefore, any time series analysis of optical data needs a previous step for gap-filling reconstruction of the time series. Numerous techniques have been presented in the literature to carry out this preliminary and mandatory step. However, the evaluation and comparison of these techniques is difficult, since no "truth" time series is available. We present here a probabilistic model (CloudSim) to provide global typical annual time series for NDVI (Normalized Difference Vegetation Index) and surface temperature (ST) time series (both over land and sea) with and without cloud contamination. This model takes advantage of the newly released Long Term Data Record Version 4 (LTDR-V4) dataset, by estimating at the pixel level daily probabilities of cloud presence, as well as mean and standard deviations for both cloud and clear acquisitions of daily NDVI and ST. Therefore, we can simulate real time series of cloud contamination influence on NDVI or ST parameters. These simulated time series allow for the assessment of the validity and usefulness of any time series reconstruction method, as well as an objective comparison of the efficiency of different methods.**

Keywords—time series reconstruction; validation; NDVI, LST, SST

I. INTRODUCTION

Cloud and atmospheric contamination of optical data is a constant and annoying feature of time series analyses, which tends to decrease artificially the values taken by parameters such as Normalized Difference Vegetation Index – NDVI [1] or surface temperatures (ST). Since cloud occurrence may reach an elevated frequency for specific areas (mountains, tropics, poles), these artificial decreases need to be identified and corrected through a gap-filling reconstruction technique for an adequate parameter time series analysis.

Numerous techniques have been presented in the literature to carry out this preliminary and mandatory step [2-10], which are described briefly for example in [10]. However, the validation of these techniques, as well as their comparison is difficult, since no "true" time series of only cloud-free observations is available. Therefore, validation and comparison between methods have been traditionally carried out through visual comparison of reconstructed time series or parameter maps [3,4,5,7,8], simulation of contaminated time series through controlled introduction of perturbations [6,10], extraction of time series metrics [2,10], or even comparison with ensemble means [11]. Only one method has been assessed through simulation of representative cloud occurrence, although at monthly time step [12].

However, all these methods are more or less subjective, since the choice of the perturbation amplitude and frequency may favor a method over another one. Moreover, such perturbations may not be representative of real cases found in time series analyses. Therefore, a global framework for time series reconstruction method validation is needed.

To that end, we developed a probabilistic model for generating representative cloud-free and cloud-contaminated time series at global scale, allowing a quantitative assessment of any time series reconstruction method, as well as fair method comparisons. For brevity considerations, examples and figures hereafter only consider NDVI parameter.

II. MATERIAL AND METHODS

A. Data

We used the recently released Long Term Data Record Version 4 (LTDR-V4 [13]) dataset, which spans more than 30 years of global daily acquisitions at 0.05º spatial resolution. We downloaded global data from July 1981 to December 2013, from which we retrieved cloud flags, and we estimated NDVI [1], Land Surface Temperature [14], and Sea Surface Temperature [14]. The two latter were summarized in a Surface Temperature (ST) parameter for each available date. A few dates (15 in total) with obvious georeferencing errors were removed from the dataset.

This work was supported by the Spanish Ministerio de Ciencia y Tecnología under Project AYA2011-29334-C02-01 (CEOS-SPAIN).

978-1-4673-7119-3/15/$31.00 ©2015 IEEE

B. *Methodology*

We designed a probabilistic model (CloudSim) to provide global typical annual time series for NDVI and surface temperature time series (both over land and sea) with and without cloud contamination.

In a first step, we identified all observations by their Day Of Year (DOY), including DOY 60 (29th February) for odd years to DOY 59 (28th February) for interannual consistency. After a preliminary analysis (not shown), we found that aggregating observations over a time window of 5 days allowed for an adequate minimum number of cloud-free observations for most pixels.

Therefore, for each DOY, each pixel, and each parameter (NDVI and ST), we aggregated all cloudy observations over a time window of 5 days (*i.e.* from DOY-2 to DOY+2), and estimated parameter average and standard deviation. The same procedure was then applied to cloud-free observations. Additionally, the probability of cloud occurrence, estimated as the percentage of cloud contaminated observations among all valid observations, was retrieved. Finally, for each pixel and each parameter, the yearly (DOY 1 to 365) cloud-free (resp. cloudy) averages and standard deviations, as well as cloud probability time series were summarized as a harmonic series with 6 modes (corresponding to periods of approximatively 12 to 2 months), resulting in 13 global maps (1 constant, 6 amplitude and 6 phase maps) for each time series and each parameter.

Cloud-free and cloudy time series for any (or all) pixels can then be created as follows:

- For each DOY, cloud-free observations are estimated from a normal random distribution with the cloud-free average and standard deviation values corresponding to this DOY.
- As previously, cloudy observations are generated for each DOY from a normal distribution with the cloudy average and standard deviation values corresponding to this DOY.

Cloud-free time series are obtained directly from the first step, while cloud-contaminated time series are generated by a combination of both steps: we first generate a uniformly random time series with values between 0 and 1, which we compare to the cloud probability (also between 0 and 1). Where the uniformly random time series is above the cloud probability, the observation is marked as cloud-free, and cloud-free observations are retained. In the other case, cloudy observations are retained (Fig. 1).

III. Application

While compiling cloud-free and cloudy observations to generate the 6-mode harmonic time series for the estimation of cloud probability and cloud-free (resp. cloudy) average and standard deviation is time consuming, the generation of the cloud contaminated and cloud-free time series from CloudSim data is straightforward. We generated such yearly time series for the whole globe for NDVI, and applied the iterative Interpolation for Data Reconstruction (IDR [10]) method as implemented in [15] to reconstruct the cloud-contaminated time series for all pixels.

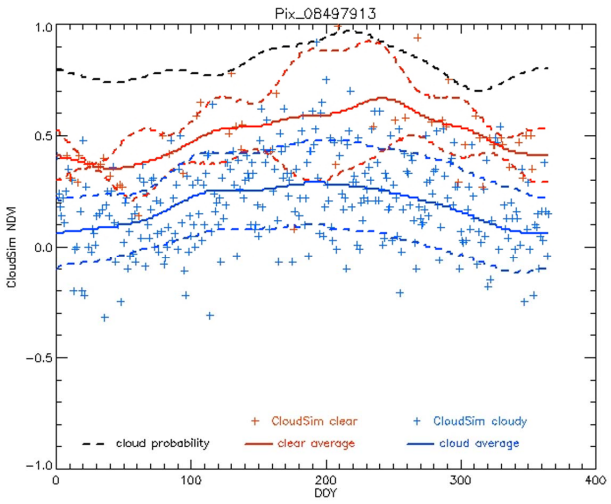

Fig. 1. Example of CloudSim cloud-free (red) and cloudy (blue) average and standard deviation (dotted line) time series for a pixel located at 31.00ºN 84.35ºW, along with the simulated CloudSim cloud contaminated time series (+, orange for cloud-free NDVI, light blue for cloudy NDVI), and cloud probability (dotted black line, between 0 and 1).

Fig. 2 maps the retrieved Root Mean Square Error (RMSE) between the generated cloud-free time series and the IDR reconstructed time series of the generated cloud contaminated time series with CloudSim. Fig. 2 shows that that the IDR method is well-suited for reconstruction of daily NDVI time series when yearly average cloud contamination is below 80% (not shown).

IV. Discussion

The methodology developed above assumes implicitly the following hypotheses are true:

- LTDR-V4 georeferenciation is perfect,
- No change occurs at pixel scale between 1981 and 2013,
- LTDR-V4 cloud flag is right, and cloudy and cloud-free observations follow a normal distribution.

As stated in the data section, georeferencing of LTDR-V4 is not perfect, 15 acquisitions have been removed from the dataset due to obvious errors, and errors of approximately one pixel are not uncommon. However, in most cases, the surrounding pixels have a similar vegetation under a similar climate, and therefore these errors should not effect the data, except for coastal areas, where errors between land and sea pixels would show easily in the data.

The no change hypothesis is also a strong one, especially in the tropics, where deforestation has been documented over the studied period [16], not to mention fire scars, land use change, or agricultural practices. However, due to the size of the observed pixel (a rough 5 km by 5 km at the Equator), these changes are averaged with no change areas, which decreases

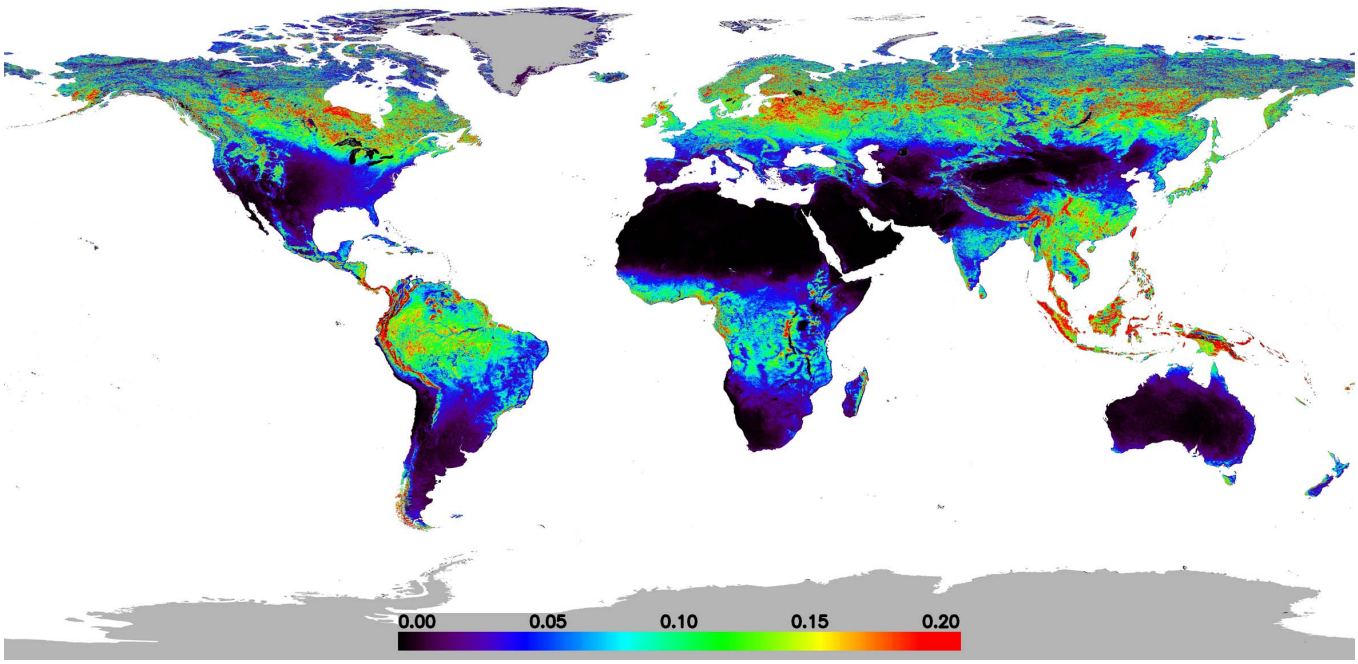

Fig. 2. Root Mean Square Error map for IDR reconstruction of a time series of one year of cloud contaminated NDVI through CloudSim.

their influence. Nonetheless, CloudSim data are to be used as simulated data for method assessment, not as a reference for land cover state.

As for the LTDR-V4 cloud flag, its validation is out of the scope of this study. Moreover, observations flagged as cloudy may still be useful for data reconstruction when parameter value is close to what it would have been under clear sky conditions. Reversely, observations flagged as cloud-free may have been effected by atmospheric aerosols and show lower than expected values. Fig. 3 presents the example of one pixel, located at 31.00ºN 84.35ºW, for which some observations labelled as cloud-free are located below most observations, while some observations labelled as cloudy are located above their cloud-free counterparts. Obviously, this breaks the widely accepted assumption that cloud effect on NDVI decreases NDVI values. Fig.3 also shows that these cases tend to increase standard deviation (dotted line), and that observation distribution within cloud-free and cloudy labels might not be considered as normal.

Therefore, georeferenciation errors and pixel changes can be ignored when using CloudSim data to validate time series reconstruction methods, although an alternative method for pixel labelling is needed for a more accurate estimation of cloud probability, as well as for average and standard deviation estimation of cloud-free and cloudy cases.

V. Conclusion

We have designed a fair benchmark for quantitative assessment of time series reconstruction technique validation and comparison, which we applied for demonstration to the IDR technique. However, LTDR-V4 cloud flag has been shown to be insufficient to estimate correctly cloud-free and cloudy time series statistics. Therefore, the quantitative assessment of IDR errors should be used as indicative of expected errors, and not as a definitive validation.

In a near future, we will develop an alternative approach to flag pixels as cloud-free or cloudy, based only on parameter values. Then, we plan to use the CloudSim model to compare quantitatively the most widely used methods for time series reconstruction, in order to provide a fair intercomparison of methods, as well as to provide users with a clear map of areas where given methods are fully reliable.

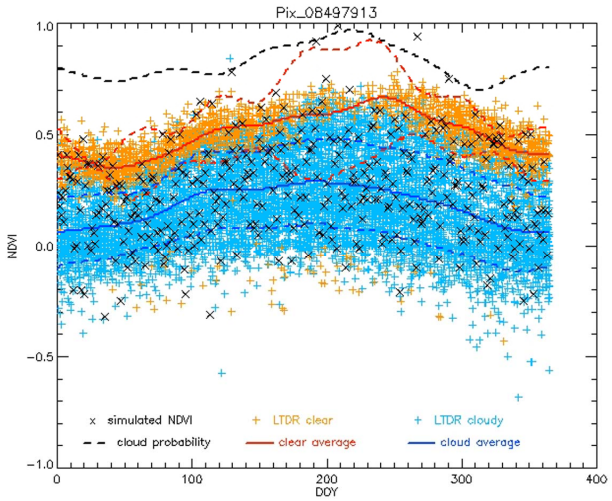

Fig. 3. Distribution of LTDR-V4 NDVI data between 1981 and 2013 for a pixel located at 31.00ºN 84.35ºW. (+) indicates observations labelled as cloud-free (orange) and cloudy (light blue), along with CloudSim time series for this pixel, as in Fig. 1, with simulated cloud contaminated time series indicated by (x).

978-1-4673-7119-3/15/$31.00 ©2015 IEEE

ACKNOWLEDGMENT

The authors want to thank NASA for providing free access to LTDR-V4 data.

REFERENCES

[1] Tucker, C. J. (1979). Red and photographic infrared linear combinations for monitoring vegetation, *Remote Sensing of Environment*, 8, 127-150.

[2] Beck, P., Atzberger, C., Hogda, K. A., Johansen, B., and Skidmore, A. (2006). Improved monitoring of vegetation dynamics at very high latitudes: A new method using MODIS NDVI, *Remote Sensing of Environment*, 100 (2006), 321-334.

[3] Chen, J., Jönsson, P., Tamura, M., Gu, Z., Matsushita, B., and Eklundh, L. (2004). A simple method for reconstructing a high-quality NDVI time-series data set based on the Savitzky–Golay filter, *Remote Sensing of Environment*, 91, 332-334.

[4] Jönsson, P., and Eklundh, L. (2002). Seasonality extraction by function-fitting to time series of satellite sensor data, *IEEE Transactions on Geoscience and Remote Sensing*, 40, 1824−1832.

[5] Jönsson, P., and Eklundh, L. (2004). TIMESAT - A program for analyzing time-series of satellite sensor data, *Computers and Geoscience*, 30, 833-845.

[6] Ma, M., and Veroustraete, F. (2006). Reconstructing pathfinder AVHRR land NDVI time-series data for the Northwest of China, *Advances in Space Research*, 37, 835-840.

[7] Roerink, G. J., Menenti, M., and Verhoef, W. (2000). Reconstructing cloudfree NDVI composites using Fourier analysis of time series, *International Journal of Remote Sensing*, 21 (No. 9), 1911-1917.

[8] van Dijk, A., Callis, S., Sakamoto, C., and Decker, W. (1987). Smoothing vegetation index profiles: An alternative method for reducing radiometric disturbance in NOAA/AVHRR data, *Photogrammetric Engineering and Remote Sensing*, 53, 1059-1067.

[9] Viovy, N., Arino, O., and Velward, A. (1992). The Best Index Slope Extraction (BISE): A method for reducing noise in NDVI time-series, *International Journal of Remote Sensing*, 13, 1585-1590.

[10] Julien, Y. and Sobrino, J. A. (2010). Comparison of cloud-reconstruction methods for time series of composite NDVI data, *Remote Sensing of Environment*, 114 (2010) 618–625.

[11] Geng, L., Ma, M., Wang, X., Yu, W., Jia, S., Wang, H. (2014). Comparison of Eight Techniques for Reconstructing Multi-Satellite Sensor Time-Series NDVI Data Sets in the Heihe River Basin, China, *Remote Sensing*, 2014, 6, 2024-2049.

[12] Zhou J., Jia, L., Hu, G. and Menenti, M. (2013). A global evaluation of harmonic analysis of time series under distinct gap conditions, *EARSeL eProceedings*, 12 (1): 58-66

[13] Pedelty, J., Devadiga, S., Masuoka, E., Brown, M., Pinzon, J., Tucker, C., et al. (2007). Generating a long-term land data record from the AVHRR and MODIS instruments, *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2007, pp. 1021–1025, doi:10.1109/IGARSS.2007.4422974.

[14] Sobrino, J. A. and Raissouni, N. (2000). Toward remote sensing methods for land cover dynamic monitoring: application to Morocco, *International Journal of Remote Sensing*, 2000, Vol. 21, No. 2, 353-366.

[15] Sobrino, J. A., Julien, Y. and Sòria, G. (2013). Estimation of land surface phenology from Meteosat Second Generation SEVIRI data (2008-2011), *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, Vol. 6, No. 3, 1653-1659.

[16] Achard, F., Beuchle, R., Mayaux, P., Stibig, H.-J., Bodart, C., Brink, A., Carboni, S., Desclée, B., Donnay, F., Eva, H. D., Lupi, A., Raši, R., Seliger, R. and Simonetti, D. (2014), Determination of tropical deforestation rates and related carbon losses from 1990 to 2010, *Global Change Biology*, 20: 2540–2554.