RECONSTRUCTION OF CLOUD-FREE TIME SERIES SATELLITE OBSERVATIONS OF LAND SURFACE TEMPERATURE

Hamid Reza Ghafarian[1], Massimo Menenti[1], Li Jia[2], and Hendrik den Ouden[1]

1. University of Delft, Department of Geoscience and Remote sensing, Delft, The Netherlands; E-mail: {H.R.Ghafarianmalamiri / M.Menenti}(at)tudelft.nl; H.denOuden(at)student.tudelft.nl
2. Alterra, Wageningen University and Research Centre, Wageningen, The Netherlands; E-mail: Li.Jia(at)wur.nl

ABSTRACT

Time series satellite observations of land surface properties, like Land Surface Temperature (*LST*), often feature missing data or data with anomalous values due to cloud coverage, malfunction of sensor, atmospheric aerosols, defective cloud masking and retrieval algorithms. Preprocessing procedures are needed to identify anomalous observations resulting in gaps and outliers and then reconstruct the time series by filling the gaps. Hourly *LST* observations, estimated from radiometric data acquired by the Single channel Visible and Infrared Spin Scan Radiometer (S-VISSR) sensor onboard the Fengyun-2C (FY-2C) Chinese geostationary satellite have been used in this study which cover the whole Tibetan Plateau from 2008 through 2010 with a 5×5 km^2 spatial resolution. Multi-channel Singular Spectrum Analysis (M-SSA), an advanced methodology of time series analysis, has been utilized to reconstruct *LST* time series. The results show that this methodology has the ability to fill the gaps and also remove the outliers (both positive and negative). To validate the methodology, we employed *LST* ground measurements and created artificial gaps. The results indicated with 63% of hourly gaps in the time series, the Mean Absolute Error (MAE) reached 2.25 Kelvin (K) with R^2 = 0.83. This study shows the ability of M-SSA that uses temporal and spatio-temporal correlation to fill the gaps to reconstruct *LST* time series.

INTRODUCTION

Land surface temperature (*LST*) is the most critical land surface property to assess the partition of available energy between sensible and latent heat flux. In many hydrological budget calculations, daily values of evaporation are needed and in some case diurnal variations of evaporation are considered. Estimation of daily evaporation is usually done with instantaneous satellite observations of land surface temperature and other land surface variables and by extrapolating the instantaneous evaporation over a day, relying on the hypothesis of a constant evaporative fraction (1,2,3).

Remote sensing time series data, which focus on the land surface properties like land surface temperature (*LST*), often features missing data and outliers, spatially and temporally. Missing data means no valid surface observations due to cloud coverage and/or malfunction of sensor. Outliers, abnormal values compared to adjacent observations in a time profile, are categorized in two different types; positive and negative. They show up either as a measured value much higher or lower than acceptable values for that property; (e.g. Normalised Differential Vegetation Indices (*NDVI*) values of more than 1) or as a measured value which is not comparable to adjacent values in time and space; (e.g. a sudden increase of a single *LST* value in a hourly sampled temporal profile). In thermal remote sensing, clouds and atmospheric aerosols usually absorb some part of the emitted thermal energy coming from the sun and the earth. They also emit thermal infrared energy as the temperature of clouds which usually is much lower than the underlying ground, so when the cloud masking algorithm does not detect clouds correctly, then we observe the temperature of a cloud, and see some negative outliers.

The geostationary satellites with their frequent observations during a day (i.e., 15 minutes to hourly observations) make surface observations more likely. As described above, the quality, spatial and temporal consistency of time series of remotely sensed *LST* data are degraded by the number,

size, distribution and continuity of gaps, so that a given point at the surface may be observable just 10% of the time. The problem will be more complicated when a data set includes both gaps and outliers. To fill the gaps and remove the outliers, a number of methods have been developed and successfully applied in some cases.

Some examples include using Fast Fourier Transform (FFT) and Harmonic Analysis of Time Series (HANTS) algorithm (4,5,6). Jia et al. (7) applied HANTS algorithm to create gap-filled ET estimated using MODIS data. Julien et al. (8) utilized HANTS for time series of yearly mean *LST* to obtain cloud-free time series and their results confirmed the usefulness of it for *LST* analysis. Even though, the HANTS algorithm has been tested for different kind of applications, still there are some limitations when it deals with large, continuous gaps. Jia et al. (9) used Temporal Similarity-Statistics (TSS) methods to find some initial values for HANTS when large continuous gaps exist in MODIS *NDVI* data using available historical data for each pixel. But, when the historical data also have gaps the TSS method does not solve the problem, because this method considers only temporal correlation between observations to fill the gaps.

In this work, we used the iterative form of Singular Spectrum Analysis (SSA) for both single channel variable (considering only one time profile) and Multi-channel (M-SSA) which consider series of time profiles. This is considered an advanced methodology which uses both temporal (SSA) and spatio-temporal (M-SSA) correlation to fill the gaps especially for continuous gaps. Singular spectrum analysis originally developed by (10) and proposed by (11,12) to be used for gap filling of time series data. In a dynamic system (e.g. diurnal variation of *LST*), individual pixel values in time, represent the outcome of the interaction among all radiative and turbulent energy exchange processes. Therefore, the evolution of whole records in time often have both regular (cycles) and irregular (noise) components. Using this idea, this method uses Empirical Orthogonal Functions (EOFs) to extract information from short and noisy time series without initial knowledge of the dynamic processes affecting the underlying time series (11). As in many time series data, a few leading components capture most variance in data sets while the rest is considered as noise. The objective of this paper is to evaluate the usefulness of SSA and M-SSA to identify gaps and remove outliers to reconstruct gap-free hourly time series satellite observation of land surface temperature. The main questions are whether it is possible: to utilize SSA and M-SSA to reconstruct diurnal variation of *LST*; to identify and fill the gaps; to identify and remove the outliers and to validate the results.

METHODS

The study area is the Tibetan Plateau which is the highest plateau in the world, located in the centre of Asia (Figure 1).

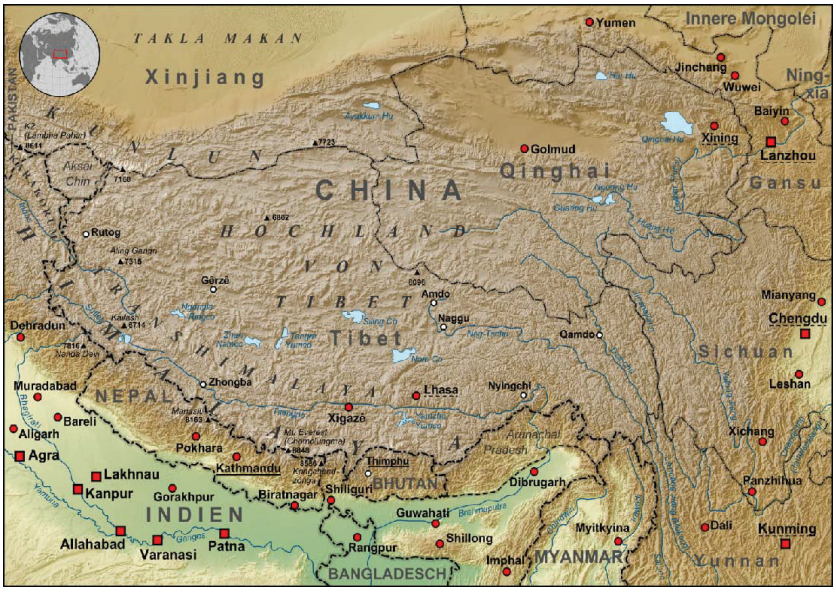


Figure 1: Tibetan plateau study area.

The Tibetan Plateau lies between the Himalaya Mountains to the south and the Taklimakan Desert to the north. Top-left corner of the study area is 39°19'39.37"N, 64°12'12.26"E and the bottom-right corner is 24°51'12.62"N, 107°2'48.11"E. It occupies an area of around 7.5 million square kilometres. It has an average elevation of 4,500 metres. In this research, land surface temperature (*LST*) extracted (13) from Single channel Visible and Infrared Spin Scan Radiometer (S-VISSR) sensor onboard the Fengyun-2C (FY-2C) spin-stabilized geosynchronous meteorological satellites (14) were used. It has one visible (VIS) channel and four infrared (IR) channels. Temporal and spatial resolution of *LST* data is hourly and 5×5 km^2, respectively. The data set covers the whole area in Figure 1 for the time period from the 1st January 2008 until 31st December 2010. There are four stations in Tibetan Plateau which measure meteorological, soil and flux data. The ground measurement data, used in this study, is from the Yingke station which measure land surface temperature every 10 minutes from 2008-2010. The Latitude and Longitude are 38°N and 100.23°E, respectively, with an elevation of 4147 metres.

The concept of multi and singular spectrum analysis is briefly described as follows.

First, we consider the univariate case which only uses temporal correlation in reconstruction of time series and later we consider the generalized form of SSA in the multivariate case which uses both spatial and temporal correlation. The whole work flow of SSA gap-filling and smoothing algorithm is illustrated as follows based on (15):

Step1: A single scalar time series $x(t)$; $t = 1, 2, \ldots, n$ is embedded into a multidimensional trajectory matrix of lagged vectors

$$\mathbf{X} = \left[x_1, \ldots, x_{n'}\right]$$

where $n' = n - m + 1$ and each lagged vector is defined as

$$\mathbf{X}_j = \left(x_j, \ldots, x_{j+m-1}\right)^T$$

This trajectory matrix contains the complete record of patterns presented within a window of size m. By selecting a large number of window size, more information about the basic pattern of the time series will be captured. A small value of window size enhances the statistical confidence at the final results (16), since the structure of time series will be captured repeatedly (17). The final form of the trajectory matrix X is a rectangular matrix of the form:

$$\mathbf{X} = \begin{pmatrix} x_1 & x_2 & x_3 & \ldots & x_{n'} \\ x_2 & x_3 & x_4 & \ldots & x_{n'+1} \\ x_3 & x_4 & x_5 & \ldots & x_{n'+3} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ x_m & x_{m+1} & x_{m+2} & \ldots & x_n \end{pmatrix}$$

Step2: The next step is decomposition of trajectory matrix **X** of size $m \times n'$ using the Singular Value Decomposition (SVD) method which yields:

$$\mathbf{X} = \mathbf{DLE}^T$$

where **D** and **E** are left and right singular vectors of **X** with $m \times m$ and $n \times n$ size, respectively, and **L** is a rectangular diagonal matrix of size $m \times n$. The elements of **L**, called singular values, are the square roots of the eigenvalues of the lagged-covariance matrix $\mathbf{S} = \mathbf{XX}^T$ of size $m \times m$. The columns of matrix **D** are the eigenvectors of **S** also known as Empirical Orthogonal Functions (EOFs). The rows of $\mathbf{E}^T$ are eigenvectors of matrix $\mathbf{X}^T\mathbf{X}$. If we plot the singular values in descending order, one can often distinguish between an initial steep slope, representing a signal, and a (more or less) flat floor, representing the noise level (11). Then any subset of d eigenvalues (EOFs), $1 \le d \le m$, for which related eigenvalues are positive provides the best representation of the matrix **X** as a sum of matrices $\mathbf{X}_i$, $i = 1,2,\ldots,d$ (18).

Step 3: Partitioning *d* eigentriples into *p* distinct subsets, then, summing all the components inside each subset such that:

$$\mathbf{X} = \sum_{n'=1}^{p} \mathbf{X}_{In'} \qquad \text{where} \qquad \mathbf{X}_{In'} = \sum_{i \in I_{n'}}^{p} \mathbf{X}_i$$

The matrices $\mathbf{X}_{In'}$ have the form of a Hankel matrix in an ideal case and consequently fit to the trajectory matrices.

Step 4: Since the ideal case described in step 3 is not usually the case, the $\mathbf{X}_{In'}$ matrices should be transformed into the form of a Hankel matrix to fit to the trajectory matrices. This step is known as diagonal averaging. In this sense, the original matrix can be reconstructed as the sum of these matrices.

$$x_t = \sum_{n'=1}^{p} x_t^{(n')}, \; t = 0, \ldots, n-1$$

Where for each *p*, the series $x_t^{(n')}$ is the result of the diagonal averaging of the matrix $\mathbf{X}_{In'}$.

The SSA workflow for gap filling procedure consists of several steps which are explained as follows:

For a given windows width (m) the original time series is centred by computing the unbiased value of the mean, and the missing data is set to zero. The first leading EOF is found by an iterative procedure which applies the SSA algorithm on the zeroed and centred set. The missing values are updated based on the reconstructed component of the current EOF. Using this updated set the SSA algorithm is applied again, and the missing values are updated based on the result of this new iteration. The process repeats until convergence is achieved. Then the iteration starts for the second leading EOF (keeping the first one fixed) until convergence has been achieved for the second EOF. This process repeats for the selected number of EOFs, each time keeping the previous ones as fixed. To find the optimal value for the window width and the number of dominant SSA modes to fill the gaps, cross-validation is applied. It means that a portion of the available data (selected as random) is flagged as missing, and the RMSE error from the reconstruction is computed to find the best value for the window size and number of EOFs.

The SSA technique can be generalized to be used for multivariate time series and gap-filling of missing values in those time series (19). We used the SSA-MTM Toolkit software in this study.[1]

RESULTS

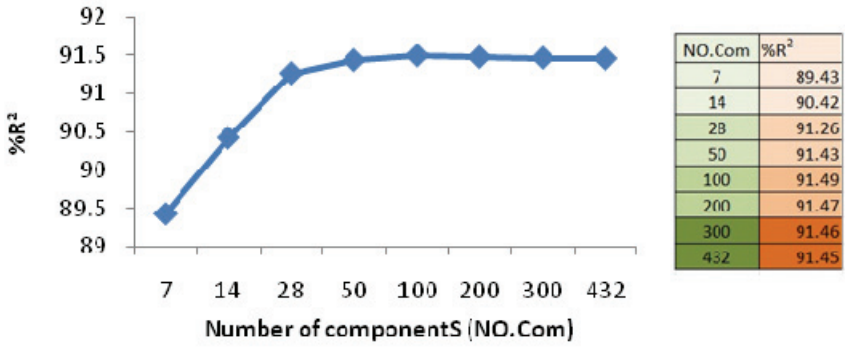
Land surface temperature time series often have a significant number of missing data and outliers. In the data set, there are different types of errors and outliers. Missing values in the time series with their different duration and location (both temporal and spatial) related to constant cloud cover and other causes are one type of problems. Another source of errors belongs to positive outliers which are due to retrieval errors. The third sources of errors belong to negative outliers due to the cloud mask algorithm which cannot detect clouds perfectly, causing the temperature of these unmasked-clouds to be measured instead of temperature of the underlying ground. Since the temperature of a cloud is always lower than the land surface temperature, this contributes to negative outliers in those situations. We divided results based on these three kinds of problems mentioned above. First we want to see whether SSA is capable of reconstructing diurnal variation of *LST*. If so, then we want to see whether SSA can remove positive and negative outliers. Finally we validate the performance of SSA using *LST* ground measurements. In the following sections, we describe results for each kind of problem separately.

[1] available for download at http://www.atmos.ucla.edu/tcd/ssa/ (last date accessed: 26 Nov 2012)

Reconstruction of *LST* time series using SSA

Here, we want to see whether it is possible to use SSA technique to reconstruct *LST* diurnal variation. As we have hourly *LST*, for each 24 hours we can see an oscillation. As described in the method section, the windows size (*m*) and main SSA dominant modes (*d*) are two main parameters for SSA. So first we want to select the optimum window size and number of relevant periodic components. To do so, we used *LST* data from the ground station measured every 10 minutes during January 2008 with a total 4464 records. This is because we need an error-free data set to see the effect of different window size and components in reconstructed time series compared to the original one. Cross-validation is then used to determine the optimum number of leading SSA main dominant modes and window size. This was done by creating artificial gaps in the ground measurements and applying SSA with different window sizes and number of components to this gappy data set. The resulting *R*-squared (R^2) values between reconstructed and original data are used to find the optimum SSA parameters based on a trade-off between a low *R*-squared value and calculation time.

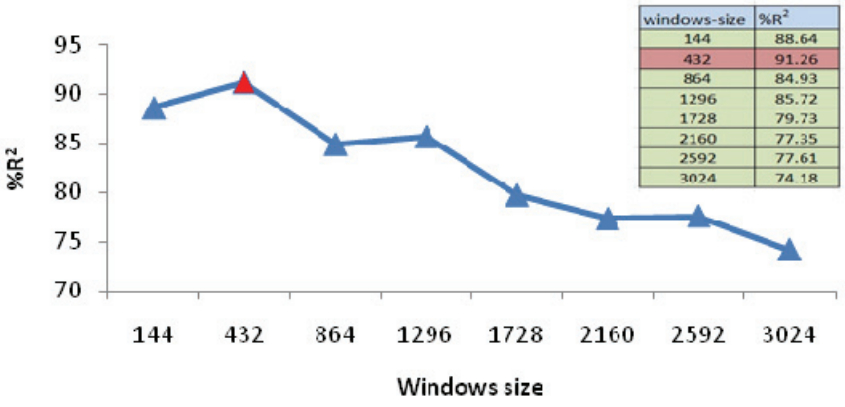
Figure 2 shows that increasing the number of components (No.com) from 7 to 28, causes the R^2 to increase from 89.43% to 91.26%. However, the calculation time for reconstruction increases considerably when going from 7 to 28 components. Since the accuracy does not change that much (just 1.83%), it was decided to use 7 as the number of component for further analysis.



NO.Com	%R²
7	89.43
14	90.42
28	91.26
50	91.43
100	91.49
200	91.47
300	91.46
432	91.45

Figure 2: Correlation coefficient of estimated and observed LST as a function of the number of components.

The same test for selecting optimum window size was conducted and as Figure 3 shows the optimum window size is 432, which is equal to 3 days or 72 hours *LST*. These values, number of components as 7 and window size as 72 hours, are now used as the main SSA parameters for reconstruction of time series satellite observation of *LST*.



windows-size	%R²
144	88.64
432	91.26
864	84.93
1296	85.72
1728	79.73
2160	77.35
2592	77.61
3024	74.18

Figure 3: Correlation coefficient of estimated and observed LST as a function of the windows size.

After selecting those parameters, we used SSA to find the ability of SSA for gap-filling of a *LST* temporal profile of the corresponding pixel of the ground stations (Figure 4). This result shows that even with 63% of gaps in the time signal, SSA was able to create a gap-free data set with the oscillations like in the original satellite data.

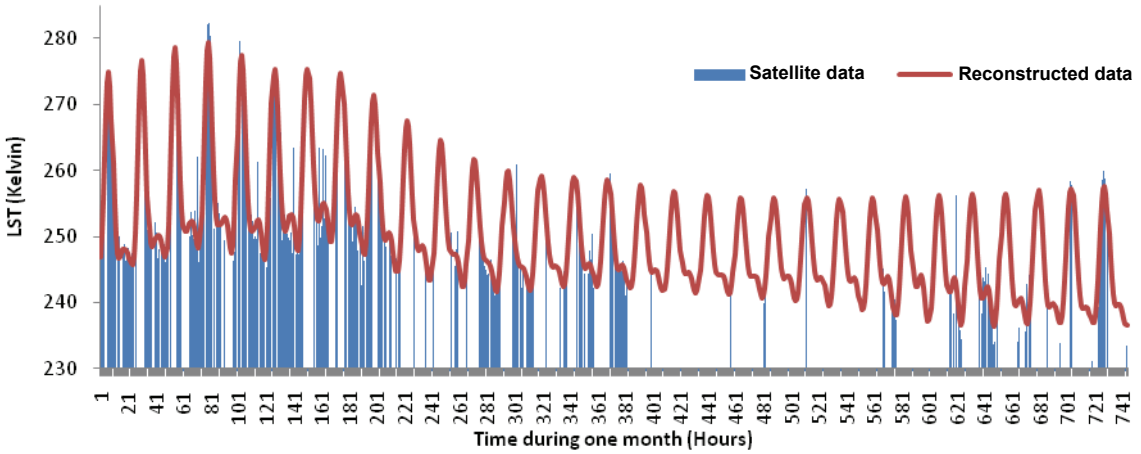


Figure 4: Gap-free reconstructed LST during January of 2008 using SSA with No.com=7 and windows size =72 hours.

Positive and negative outliers removal

The existence of outliers (negative and positive) as well as gaps imposes additional challenges to the reconstruction of *LST* time series. In many time series data sets which have periodic components like diurnal variation of *LST*, the broad, slow variations that offer some degree of periodicity (signals) are of greater interest than the fast changes, which often appear as random, unpredictable events (noise), such as uncertainties in the observations (like outliers in our case) (15). When a small number of components which belong to the signal have been selected, the remaining noise, which belongs to outliers and other sources of error, has been removed (Figure 5), but still the effect of continuing to include them in calculation causes some deviation from the original data set. In order to remove these effects, we first applied M-SSA to the original time series of *LST* and estimated the reconstructed time series (Figure 5, red line) with predefined M-SSA parameters. Then the absolute differences between the original time series values and first reconstructed result were calculated. As explained before, negative outliers are mainly observed as the temperature of clouds. Cloud's temperature varies in terms of their altitude, type and thickness. Since these clouds are regularly composed above the earth surface (i.e., above 1 km), it is expected to present lower temperature than the background surface (nearly 1 degree per 100 metres altitude). Therefore, a value of 10 Kelvin was defined as the threshold between the deviation of the original observations and the reconstructed data. This sets deviations of more than 10 Kelvin to the value zero in the original observations. Then we again apply M-SSA on the new time series in which the outliers have been removed. The results show that the reconstructed values after removing the outliers become more similar to the original data than the reconstructed values before removing the outliers (Figure 6, red line).

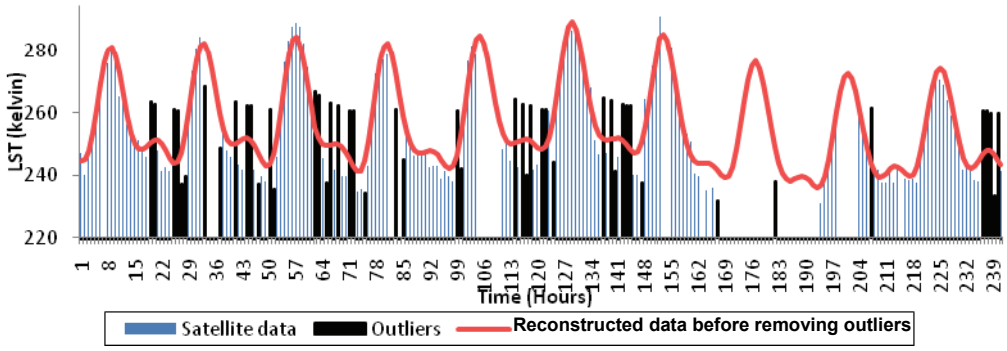


Figure 5: Reconstruction and outliers removal using M-SSA.

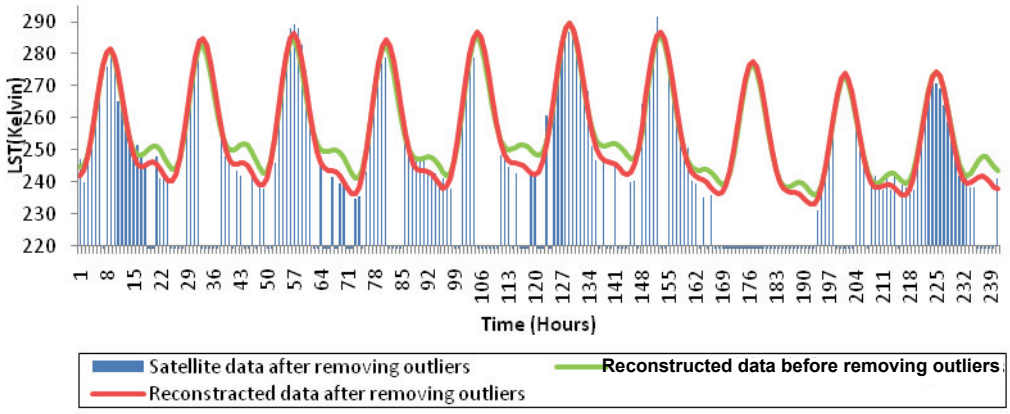


Figure 6: Reconstruction of LST before and after outliers removal.

Validation of SSA using ground measurement *LST*

In order to validate the results, we use cross-validation method. In cross-validation, we use data set itself and create some randomly artificial gaps in that time series and then compare the results of reconstruction with original data set. In this case, it is better having a gap-free original data set. As original satellite data contain error and noise, cross-validation of results, by creating some artificial gaps, is not reliable. In order to validate results of SSA to reconstruct time series *LST*, we used time series ground measurements of *LST* which were measured in Yinkle station in the same time periods (January 2008) as satellite observations. Because we now have actual *LST* measurements, we can use them to validate the performance of SSA in gap-filling. To do so, we should create some randomly artificial gaps on them. As the time period of both ground and satellite observation is the same and also in order to have randomly distributed gaps, we used the gaps pattern of covered ground pixel and imposed the same gaps pattern on ground measurements. Then SSA was applied on gappy ground measurements, and the result was compared with actual data. In Figure 7, the red lines show the gaps while the blue lines belong to actual ground measurements. The black line shows the results of applying SSA to filling the gaps. From the above results, we found that even with 63% of gaps, the R^2 = 0.83 with MAE = 2.25 Kelvin.

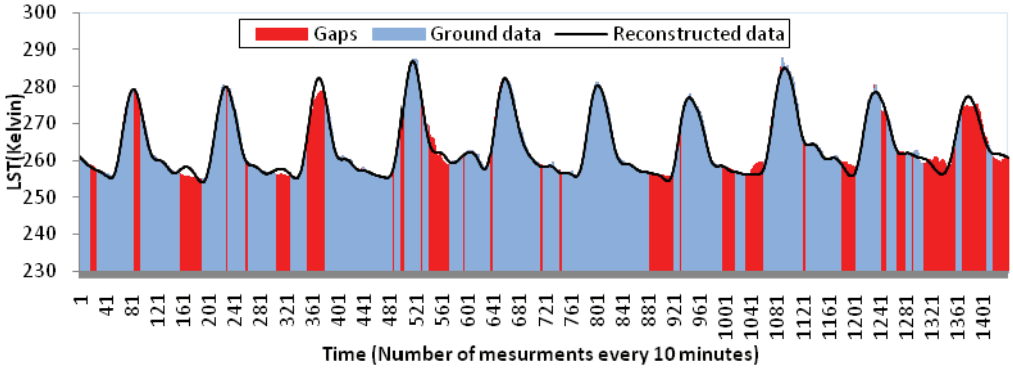


Figure 7: Validation of SSA gap-filling using ground measurements with the same pattern of gaps as overlying pixels from satellite.

CONCLUSION

SSA gap-filling method was tested for hourly time series of *LST*. As the actual time series has a lot of continuous gaps, with the help of this method, results with some reasonable accuracy can be obtained. The results show even with extremely complicated situations related to gaps and outliers, SSA have the ability to reconstruct gap-free data sets. It is recommended that the performance of SSA is tested with synthetic *LST* data having different number, size, continuity and location of gaps during the same time periods.

ACKNOWLEDGEMENTS

This work is jointly supported by the EU-FP7 project CEOPAEGIS (grant number 212921) and the project *Sustainable Spatial Development of Ecosystems, Landscapes, Seas and Regions* (KB-14-001-029) funded by the Ministry of Economic Affairs, The Netherlands.

REFERENCES

1 Jackson R D, J L Hatfield, R J Reginato, S B Idso & P J Pinter Jr, 1983. Estimation of daily evapotranspiration from one time-of-day measurements. Agricultural Water Management, 7(1-3): 351-362

2 Shuttleworth W J, R J Gurney, A Y Hsu & J P Omrsby, 1989. FIFE: the variation in energy partition at surface flux sites. IAHS Publications, 186: 67-74

3 Nichols W E & R H Cuenca, 1993. Evaluation of the evaporative fraction for parameterization of the surface energy balance. Water Resources Research, 29(11): 3681-3690

4 Menenti M,S Azzali,W Verhoef & R van Swol, 1993. Mapping agroecological zones and time lag in vegetation growth by means of fourier analysis of time series of NDVI images. Advances in Space Research, 13(5): 233-237

5 Verhoef W, M Menenti & S Azzali, 1996. Cover A colour composite of NOAA-AVHRR-NDVI based on time series analysis (1981-1992). International Journal of Remote Sensing, 17(2): 231-235

6 Roerink G J & M Menenti, 2000. Reconstructing cloudfree NDVI composites using Fourier analysis of time series. International Journal of Remote Sensing, 21(9): 1911-1917

7 Jia L, G Xi, S Liu, C Huang, Y Yan, & G Liu, 2009. Regional estimation of daily to annual regional evapotranspiration with MODIS data in the Yellow River Delta wetland. Hydrology and Earth System Sciences, 13: 1775-1787

8 Julien Y, J.A Sobrino & W Verhoef, 1999. Changes in land surface temperatures and NDVI values over Europe between 1982 and 1999. Remote Sensing of Environment, 103(1): 43-55

9 Jia L, H Shang, G Hu & M Menenti, 2011. Phenological response of vegetation to upstream river flow in the Heihe Rive basin by time series analysis of MODIS data. Hydrology and Earth System Sciences, 15: 1047-1064

10 Broomhead D S & G P King, 1986. Extracting qualitative dynamics from experimental data. Physica D: Nonlinear Phenomena, 20(2-3): 217-236

11 Vautard R, P Yiou & M Ghil, 1992. Singular-spectrum analysis: A toolkit for short, noisy chaotic signals. Physica D: Nonlinear Phenomena, 58(1-4): 95-126

12 Kondrashov D & M Ghil, 2006. Spatio-temporal filling of missing points in geophysical data sets. Nonlinear Processes in Geophysics, 13: 151-159

13 Tang B, Y Bi, Z L Li & J Xia, 2008. Generalized split-window algorithm for estimate of land surface temperature from Chinese Geostationary FengYun Meteorological Satellite (FY-2C) data. Sensors, 8(2): 933-951

14 NSMC, 2012. National Satellite Meteorological Center. http://www.nsmc.cma.gov.cn/newsite/NSMC_EN/Home/Index.html (last date accessed: 23 Feb 2012)

15 Musial J P, M M Verstraete & N Gobron, 2011. Technical Note: Comparing the effectiveness of recent algorithms to fill and smooth incomplete and noisy time series. Atmospheric Chemistry and Physics, 11: 7905-7923

16 Elsner J B & A A Tsonis, 1996. Singular Spectrum Analysis: A New Tool in Time Series Analysis (Plenum Press: New York) 161 pp.

17 Allen M R, M D Dettinger, K Ide, D Kondrashov, M E Mann, A W Robertson, A Saunders, Y Tian, F Varadi & P Yiou, 2002. Advanced spectral methods for climatic time series. Reviews of Geophysics, 40(1): 3.1-3.41

18 Golyandina N, V Nekrutkin & A Zhigljavsky, 2001. Analysis of Time Series Structure: SSA and Related Techniques (Chapman & Hall/CRC, Washington DC, USA) 310 pp.

19 Schoellhamer D.H, 2001. Singular spectrum analysis for time series with missing data. Geophysical Research Letters, 28(16): 3187-3190