A Multiple-point Geostatistics Method for filling gaps in Landsat ETM+ SLC-off images

Gaohong Yin[a], Gregoire Mariethoz[b] and Matthew F McCabe[a]

[a] *Division of Biological and Environmental Sciences and Engineering, King Abdullah University of Science and Technology, Saudi Arabia*

[b] *Institute of Earth Surface Dynamics, University of Lausanne, 1015 Lausanne, Switzerland*

Email: gaohong.yin@kaust.edu.sa

Abstract: The Scan Line Corrector (SLC), which compensates for the forward motion of Landsat 7, failed on May 31, 2003. The lack of SLC resulted in data gaps in the observed images, affecting the spatially continuous fields that were usually provided. Fortunately, the observations acquired by Landsat 7 are highly geometrically and radiometrically accurate compared with many other sensors, allowing the opportunity to develop gap-filling approaches to address the problem. How best to do this remains an open question. A variety of simple and effective methods based on different ideas have been proposed. Most of these can be classified into two categories: deterministic interpolation or geostatistical estimation. Deterministic interpolation approaches calculate the values of the unknown pixels based on an assumed continuity with the values of neighbouring data points. Geostatistical approaches formulate a spatial model of variability that is used to estimate the missing values as well as to quantify the uncertainty of these missing values. Most current approaches are only applicable for relatively homogeneous areas. For example they cannot satisfactorily predict the presence of narrow or small objects such as roads or streams. As a result, such objects can be truncated after interpolation. This problem may drastically constrain the application of filled SLC-off images. In this study, a multiple-point geostatistics approach, the Direct Sampling method, is adopted as a solution to fill Landsat 7 images. The Direct Sampling method uses a conditional stochastic resampling of known areas in the observation image to simulate the unknown locations. This approach can reuse the complex patterns present in the incomplete image, and has the capacity to simulate narrow or small objects. Moreover, being a geostatistical method, it allows for the generation of multiple interpolations to compute uncertainty bounds on the interpolated values. Here, the Direct Sampling method is applied to both univariate and multivariate cases to demonstrate its application. Numerical experiments indicate that the Direct Sampling method is able to fill the gaps satisfactory, especially when combining the target image with a temporally close image. The results are satisfactory for filling narrow objects like roads or streams. Compared with other gap filling method, the Direct Sampling method is relatively simple and easily employed. However, it also has limitations, such as the appropriate selection of parameters, which can greatly influence the simulation effect and computation time. Further research advancing the optimisation and providing guidance on parameter selection is required.

Keywords: *Landsat ETM+, Gap filling, Multi-point geostatistics, Direct Sampling*

1. INTRODUCTION

The Landsat Program is a joint effort between the U.S. Geological Survey (USGS) and the National Aeronautics and Space Administration (NASA), which aims to deliver a long-term record of natural and human-induced changes in the global landscape (USGS 2012). Among a fleet of current satellite missions, Landsat 7 is outstanding because of its capability to acquire highly geometrically and radiometrically accurate data worldwide (USGS 2012). Landsat data have been used in a variety of applications in the Earth and environmental sciences, including the estimation of evaporation (Ershadi et al, 2013), snow cover (Bormann et al. 2012) and vegetation retrieval (Houborg et al., 2015) to name a few. However, in May 2003, unusual artifacts appeared in collected data due to a failure in the Scan Line Corrector (SLC) instrument on the Enhanced Thematic Mapper Plus (ETM+). With this failure, the line of sight traces a zigzag pattern, resulting in data gaps in the observed images.

Considering the significance of data collected by Landsat 7 for studies of the Earth system, many methods have been proposed to fill the gaps that now exist in ETM+ images. This problem is more general since gaps also occur in other sensors as a result of orbital characteristics or meteorological events (Mariethoz et al, 2012). Remote sensing gap-filling approaches can be broadly classified into either deterministic interpolation or geostatistical stochastic based approaches. Deterministic interpolation is a family of methods that use mathematical functions to calculate the values at unknown locations, based either on the degree of similarity or the degree of smoothing in relation with neighbouring data points (Peralvo and Maidment 2003). Representative deterministic interpolation techniques include the local linear histogram matching technique provided by the USGS (2004), the multi-scale segmentation approach (Maxwell et al. 2007) and the Neighbourhood Similar Pixel Interpolator (NSPI) method (Chen et al. 2011). Geostatistical approaches make stochastic predictions and are therefore usually able to estimate values at unknown locations and provide an uncertainty range. Many geostatistical methods have been applied to fill Landsat gaps, such as kriging, co-kriging, and the Geostatistical Neighbourhood Similar Pixel Interpolator (GNSPI) (Pringe et al. 2009; Zhang et al. 2007; Zhu et al., 2012).

Among the many different approaches, one common issue is that most are only applicable to relatively homogeneous regions, while for heterogeneous areas, the quality of simulated results can be quite variable. This is especially true when estimating small features or narrow objects such as roads and streams, with discontinuities or even truncation of objects occurring. Another issue relates to the computational time cost. The computing speed of kriging or co-kriging methods can be very slow, which limits their utility as efficient solvers. Moreover, for deterministic methods, the simulation results are highly influenced by the quality of input images, which means that when the input images are not taken in identical conditions as the target image, the gap filling results can be very poor. For traditional geostatistical methods, results are generally only plausible under the intrinsically stationary assumption, which cannot be satisfied when land cover changes are a spatially and temporally varying process.

This paper explores the use of the Direct Sampling multiple-point statistics (MPS) method (Mariethoz et al. 2010) to fill the SLC gaps. MPS is a stochastic method allowing for the generation of multiple reconstructions and allows for an analysis of resulting uncertainty. It has been used in a variety of applications, including the estimation of remote sensing data based on secondary attributes (Mariethoz et al, 2009) and downscaling of climate simulations (Jha et al, 2013). Here we explore the use of multi-temporal training images, with experiments showing that the Direct Sampling method performs satisfactorily, without demanding high quality input images.

2. DESCRIPTION OF THE DIRECT SAMPLING MPS APPROACH

The basic idea of the Direct Sampling method is to use a training image to identify spatial features and properties that can be used to fill the gaps. Hereon, the image with gaps to be filled is referred to as the target image, while the images that provide information for filling gaps in the target image is identified as the input image. There are two main ways to use input imagery for the Direct Sampling method, which are defined as i) the training image (i.e. provide the required spatial and temporal information) and ii) the auxiliary image (i.e. an auxiliary variable to provide additional information when simulating). When a large portion of the target image is already known, it is possible to infer information without use of a training image, by just using the non-gap regions of the target image to inform the sampling process.

For instance, let $Z(x)$ be the variable concerned, where x is a pixel in the gaps of the target image. The aim of the Direct Sampling method is to find one possible outcome of Z conditioned to N_x from the conditional cumulative function as shown in Equation (1), where N_x is the ensemble of the n closest pixels of x that are informed (Mariethoz et al. 2012):

$$F(z) = Prob(Z(x) \leq z|N_x) \quad (1)$$

The concept behind the Direct Sampling method is to find another pixel y in the training image with a set of neighbour pixels N_y that is similar to N_x. These *n* pixels are called the neighbourhood. To compare the similarity between N_x and N_y, the concept of distance *d* (N_x, N_y) is used. A distance is a flexible concept that can be adapted to both categorical and continuous attributes. In this paper we use a Weighted Euclidian distance:

$$d(N_x, N_y) = \frac{1}{\eta}\sqrt{\sum_{i=1}^{n} w_i[Z(x_i) - Z(y_i)]^2} \quad (2)$$

where w is the weight of each point and η is a normalisation factor ensuring that the distance values remain bounded within the interval [0, 1].

The search for *y* in the training image is done in a random manner. The first time a pixel *y* is found in the training image that results in a distance below a predefined threshold *t*, the value *Z*(*y*) is assigned to *Z*(*x*). If the search area has reached a fraction *f* of the whole training image without finding a *y* that can satisfy the threshold requirement, the lowest distance is then accepted and assigned to *Z*(*x*). The process is illustrated in Fig 1. The data event is defined in Fig 1a: the central pixel with a question mark is the target pixel to be filled, and the two white and the black pixels are neighbourhoods whose values are already known. Here it is a categorical case with values contains only two options – black and white. Using the defined data event to search in the search window of the training image. If the values of neighbourhoods are not the same as data event (Fig1 c), it will go to the next location until the simulation data event is satisfactorily matched, as shown in Fig 1d, and then the value of the central pixel first matching the data event is assigned to the target pixel.

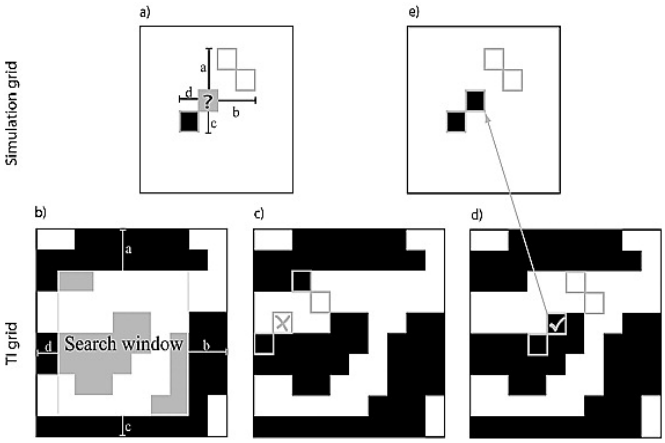

Figure 1. Illustration of DS method. (a) Define the data event.(b) Define a Search Window in the TI. (c) Search in the TI until (d) The simulation data event can be satisfactorily matched. (e) The value of central pixel first matching data event is assigned to the target pixel (Mariethoz et al., 2010)

The Sampling method can also be used for the multivariate case. When several variables need to be reconstructed, the simulation process is almost the same and the distance becomes a weighted average of the distance taken individually for each univariate neighbourhood:

$$d(N'_x, N'_y) = \sum_{j=1}^{m} \frac{\alpha_j}{\eta_j}\sqrt{\sum_{i=1}^{n} w_i^k[Z^k(x_i) - Z^k(y_i)]^2} \quad (3)$$

where *m* is the number of variables and η_k is the normalisation constants for each variable. A more detailed description of the Direct Sampling method is provided in the study of Mariethoz et al. (2010).

3. APPLICATION OF THE MPS APPROACH TO GAP-FILLING LANDSAT DATA

Our study area focused on a region in California which is covered by World Reference System 2 Path 43 and Row 34. A 401×501 sub-region (approximate 12km×15km) was selected to validate the application of the Direct Sampling method to univariate simulation and bivariate simulation cases. The study area is chosen to ensure both homogeneous (e.g. dense farmland) and heterogeneous (e.g. streams) landscape features and can be considered as a representative example to demonstrate the performance of the Direct Sampling method. In this application, only Landsat band 4 data are considered, although any number of bands could be employed.

The target image used is a Landsat 7 ETM+ image without gaps acquired on July 24, 2002 (Figure 2a). Gaps are created manually using the gap mask from an actual SLC-off ETM+ image, as shown in Figure 2b. For input images, we consider retrievals over the same area acquired on March 2, 2002 (Figure 3a), July 8, 2002 (Figure 3b) and August 25, 2002 (Figure 3c) respectively. Before simulations are performed, all input images are geometrically rectified to match the target image, and for all images, digital numbers are calibrated to top-of-atmosphere reflectance (Zhu et al. 2012).

There are 6 different cases that will be considered. Cases 1-3 use the input image of March 2, July 8 and August 25 as training images separately, which requires searching and identifying a value in the training image to fill the pixel in the gap. Case 4 does not require an input image and only uses the non-gap locations of the target image itself to fill the gaps. Cases 5 and 6 consider bivariate simulations, taking the reflectance value of the target image as the first variable, while another input image functions as an auxiliary image, with its reflectance value serving as the auxiliary variable. The dates for the auxiliary data are March 2 and July 8.

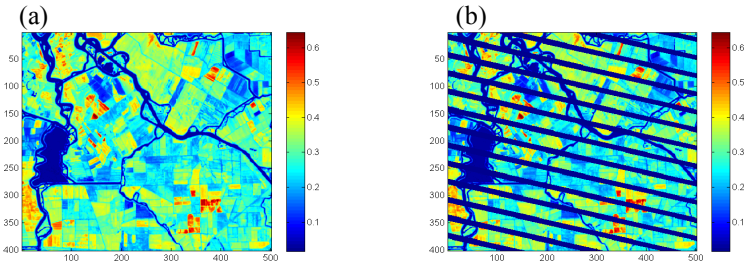

Figure 2. Studied image acquired on July 24, 2002, with (a) the actual image and (b) gaps generated manually following application of the gap mask

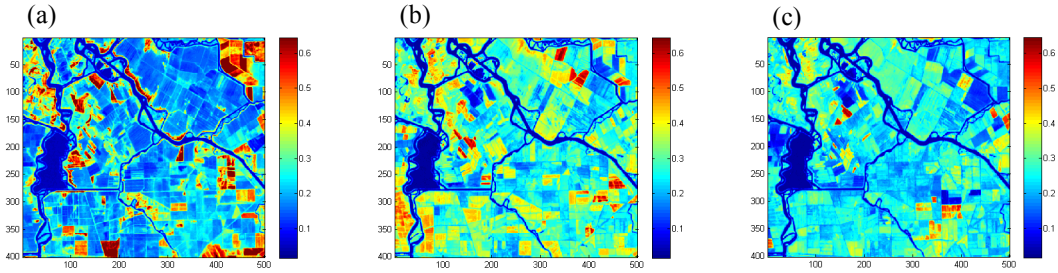

Figure 3. Input images with (a) Landsat data acquired on March 2, 2002 and data acquired on (b) July 8 and (c) August 25, 2002.

To evaluate the resulting gap-filled images, both qualitative and quantitative measures are considered. From a qualitative perspective, the existence of obvious artifacts or discontinuity can be observed, and the prediction error at each pixel compared against the true image. From a quantitative viewpoint, RMSE is an accepted measure of the general performance and is regularly employed in similar simulation studies. However, the RMSE is sensitive to occasional large errors, making it inappropriate as the only performance metric. QQ-plots, scatter plots and error distributions are also used to evaluate results. Consideration of all of these criterion allows a judgement on the simulation accuracy. Here, only a selection of the measures are displayed.

4. RESULTS

The mean of the 10 reconstructions for each of the 6 test cases is shown in Figure 4. Obvious artifacts are apparent in the realisations of Case 1 to Case 3, where the trace of previous gaps is clear. However, the last three cases represent more integrated images, although discontinuities remain in a few locations of Case 4, which only uses the gap-free information as training data. To better compare the simulation results of Case 5 and Case 6, a small part of the scene is enlarged and shown in Figure 5. When compared with the original image, Case 5, which combines a date furthest from the target date, shows blurred regions in some places, especially for narrow rivers. Case 6, which uses as covariate a temporal image closer to the target, shows much better connectivity characteristics.

Error statistics concur with the qualitative assessment, with Case 6 providing improved results, as shown in Table 1. The root mean square error (RMSE) of Case 6 is much smaller than all other cases, especially when compared with Case 1-3. The RMSE of Case 6 is over 50% lower than Case 1-3. Fig 6 displays the histograms of simulation errors, which uses reconstructed values minus the true values. The errors for all six cases are mostly unbiased, with Case 4 to 6 showing steeper and more symmetric distributions. For Case 1 to 4, most errors in reflectance values are within the range [-0.2, 0.2], while the corresponding range for Case 5 and Case 6 is [-0.1, 0.1], with Case 6 having a narrower range than Case 5. From the quantitative perspective,

it is also demonstrated that the bivariate simulation offers a better prediction when compared with other conditions, especially when the auxiliary image shows similar reflectance characteristics to the target image.

Table 1. RMSE of all six cases

	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6
RMSE	0.0903	0.0754	0.0750	0.0658	0.0497	0.0378

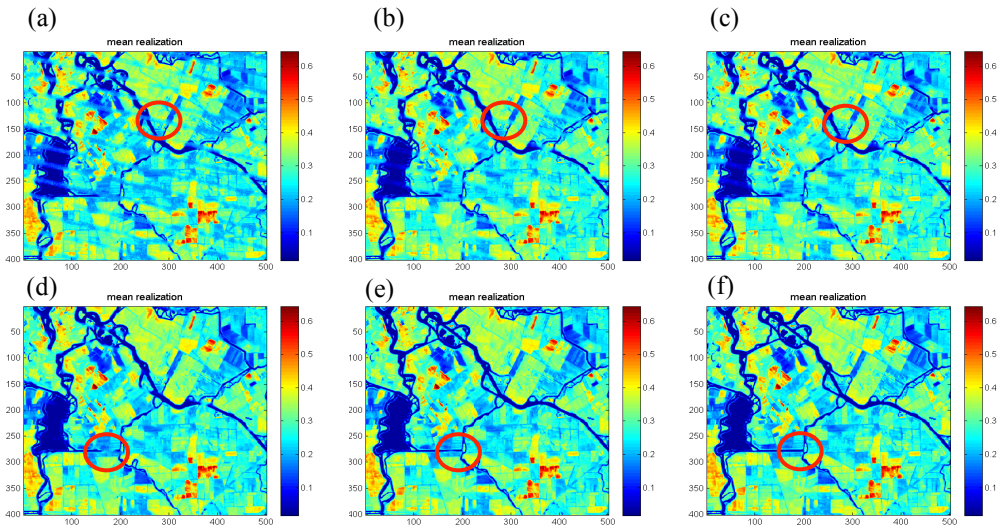

Figure 4. Mean realisations of 6 test cases, with (a)-(c) correspond to Case 1 to 3, and (d)-(f) corresponds to Cases 4, 5 and 6.

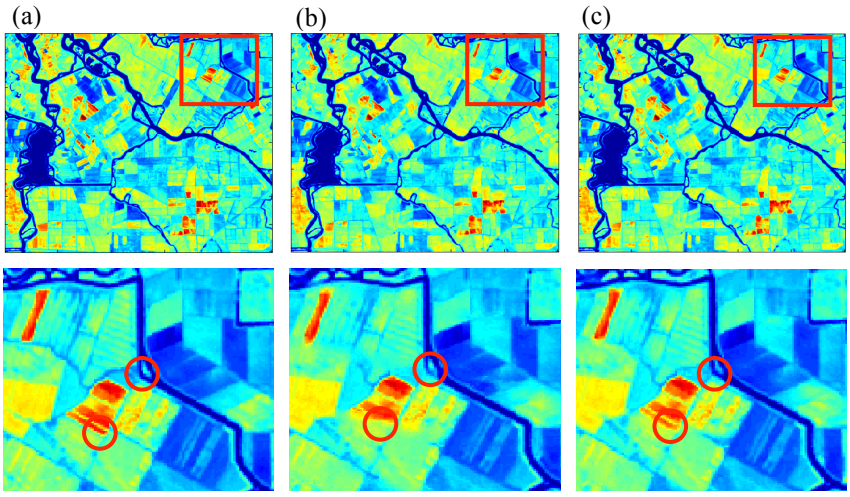

Figure 5. Detail comparison of Case 5 and Case 6, with (a) the original image of July 24, 2002 and (b) Case 5, where the target image combines the overpass of March 2, 2002 and (c) Case 6, where the target image combines the overpass with July 8, 2002.

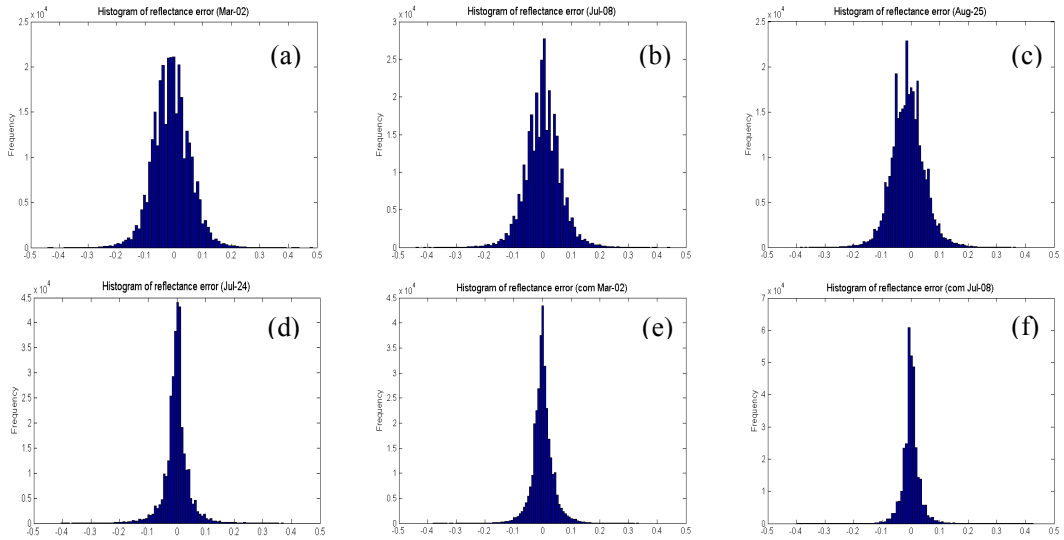

Figure 6. Error histograms for all six cases, with (a)-(f) corresponds to Case 1 to 6.

5. DISCUSSION AND CONCLUSIONS

Many methods have been proposed to fill gaps following the SLC failure. Here, we implement a newly developed geostatistical approach, the Direct Sampling method, to fill these gaps. The Direct Sampling method is a conditional stochastic approach that is based on training images. It has a prominent advantage in that it is a straightforward approach to implement and does not have strict requirements on input images when compared with traditional geostatistical methods such as kriging. Even when using the information contained within a single image (without any input images), it has been shown to provide acceptable results. If an auxiliary image that covers the same area on an alternative date is combined, it allows for further improvement in the gap filling results. Understanding the impact of temporal separation between the target and input imagery is the subject of ongoing research. Because of the complementarity between images, small or narrow objects can be predicted more accurately when auxiliary images are used.

An advantage of the MPS approach is the improvement in computational speed. To fill a target image with around 50,000 pixels using a standard Windows system with an Intel Core i7 2.80 GHz processor and 16GB of RAM, takes only a few seconds for Case 1-3, around 5 minutes for Case 4 and 20 minutes for Case 5 and 6 for each reconstruction. For comparison, the time needed to fill 5,000 missing pixels under a Linux operating system with a Xeon 2.33 GHz processor and 48GB of RAM using ordinary co-kriging, is approximately 12 hours (Pringle et al., 2009).

There are numerous applications of this approach beyond just gap-filling Landsat data. In remote sensing, gaps can result as a function of orbital characteristics and cloud cover, and effect the production of spatially consistent retrievals (McCabe et al. 2005; Liu et al. 2012) This is especially true for applications operating at regional or global scales, where spatio-temporal gaps limit the utility of such products for model evaluation (McCabe et al. 2005; Stisen et al. 2011) or spatial assessment (Manfreda et al. 2007). Developing new techniques to address such problems are certainly needed. However, one challenge for applying this method is the selection of parameters. There are three main parameters that need to be considered, including the threshold, fraction and the number of neighbourhoods. The characteristics of target image and input images effect the selection of parameters significantly. The basic principle is that a small threshold t, a large fraction f and large number of neighbourhood n give better results. However, this comes at the cost of computation time. Furthermore, if the threshold is too small and the fraction and number of neighbourhood too large, it is possible that no pixel can fulfill these conditions and the lowest distance value is given to the pixel in gaps. When too many unknown pixels are present (i.e. very large gaps), there might be very few patterns that can be resampled and the results can be unreliable. Therefore, a balance between accuracy and computing time needs to be considered. Further research is required to provide guidance on the selection of parameter values. The application of the DS method to a variety of land cover types ranging from homogeneous landscapes (e.g. desert and sparse agricultural area) to heterogeneous areas (e.g. dense farmland and urban area) is currently under investigation, and the accuracy of the DS method across diverse land types needs to be further studied.

ACKNOWLEDGMENT

Research reported in this publication was supported by the King Abdullah University of Science and Technology (KAUST). We would like to thank the USGS Landsat Science Team for providing ETM+ data.

REFERENCES

Bormann, K. J., McCabe, M. F., & Evans, J. P. (2012). Satellite based observations for seasonal snow cover detection and characterisation in Australia. Remote Sensing of Environment, 123, 57-71.

Chen, J., Zhu, X., Vogelmann, J.E., Gao, F. and Jin, S. (2011) A simple and effective method for filling gaps in Landsat ETM+ SLC-off images. Remote Sensing of Environment, 115(4), 1053-1064.

Ershadi, A., McCabe, M. F., Evans, J. P., & Walker, J. P. (2013). Effects of spatial aggregation on the multi-scale estimation of evapotranspiration. Remote Sensing of Environment, 131, 51-62.

Houborg, R., McCabe, M., Cescatti, A., Gao, F., Schull, M., & Gitelson, A. (2015). Joint leaf chlorophyll content and leaf area index retrieval from Landsat data using a regularized model inversion system (REGFLEC). Remote Sensing of Environment, 159, 203-221.

Jha, S. K., Mariethoz, G., Evans, J. P., & McCabe, M. F. (2013). Demonstration of a geostatistical approach to physically consistent downscaling of climate modeling simulations. Water Resources Research, 49(1), 245-259.

Liu, Y. Y., de Jeu, R. A., McCabe, M. F., Evans, J. P., & van Dijk, A. I. (2011). Global long-term passive microwave satellite-based retrievals of vegetation optical depth. Geophysical Research Letters, 38(18).

Manfreda, S., McCabe, M. F., Fiorentino, M., Rodríguez-Iturbe, I., & Wood, E. F. (2007). Scaling characteristics of spatial patterns of soil moisture from distributed modelling. Advances in water resources, 30(10), 2145-2150.

Mariethoz, G., P. Renard and R. Froidevaux (2009). Integrating collocated auxiliary parameters in geostatistical simulations using joint probability distributions and probability aggregation. Water Resources Research 45(8).

Mariethoz, G., Renard, P. and Straubhaar, J. (2010) The direct sampling method to perform multiple-point geostatistical simulations. Water Resources Research, 46(11).

Mariethoz, G., McCabe, M.F. and Renard, P. (2012) Spatiotemporal reconstruction of gaps in multivariate fields using the direct sampling approach. Water Resources Research, 48(10).

Maxwell, S.K., Schmidt, G.L. and Storey, J.C. (2007) A multi-scale segmentation approach to filling gaps in Landsat ETM+ SLC-off images. International Journal of Remote Sensing, 28(23), 5339-5356.

McCabe, M. F., Kalma, J. D., & Franks, S. W. (2005). Spatial and temporal patterns of land surface fluxes from remotely sensed surface temperatures within an uncertainty modelling framework. Hydrology and Earth System Sciences Discussions, 9(5), 467-480.

McCabe, M. F., Franks, S. W., & Kalma, J. D. (2005). Calibration of a land surface model using multiple data sets. Journal of Hydrology, 302(1), 209-222.

Peralvo, M. and Maidment, D. (2003) Influence of DEM interpolation methods in drainage analysis. GIS Hydro, 4.

Pringle, M.J., Schmidt, M. and Muir, J.S. (2009), Geostatistical interpolation of SLC-off Landsat ETM+ images. ISPRS Journal of Photogrammetry and Remote Sensing 64(6), 654-664.

Stisen, S., McCabe, M. F., Refsgaard, J. C., Lerer, S., & Butts, M. B. (2011). Model parameter analysis using remotely sensed pattern information in a multi-constraint framework. Journal of Hydrology, 409(1), 337-349.

USGS (2012) Landsat-A Global Land-Imaging Mission.

USGS (2004) Phase 2 gap-fill algorithm: SLC-off gap-filled products gap-fill algorithm methodology.

Zhang, C., Li, W., & Travis, D. (2007). Gaps‐fill of SLC‐off Landsat ETM+ satellite image using a geostatistical approach. International Journal of Remote Sensing, 28(22), 5103-5122.

Zhu, X., Liu, D. and Chen, J. (2012) A new geostatistical approach for filling gaps in Landsat ETM+ SLC-off images. Remote Sensing of Environment, 124, 49-60.