Environmental Modelling & Software 30 (2012) 139–142

Contents lists available at SciVerse ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft

Short communication

A three-dimensional gap filling method for large geophysical datasets: Application to global satellite soil moisture observations

Guojie Wang [a,*], Damien Garcia [b], Yi Liu [c,d], Richard de Jeu [a], A. Johannes Dolman [a]

[a] Department of Earth Sciences, VU University Amsterdam, 1085 HV Amsterdam, The Netherlands
[b] CRCHUM-Research Center, University of Montreal Hospital, Montreal, Canada
[c] Climate Change Research Centre, University of New South Wales, Sydney, Australia
[d] CSIRO Land and Water, Black Mountain Laboratories, Canberra, Australia

ARTICLE INFO

Article history:
Received 31 January 2011
Received in revised form
31 August 2011
Accepted 24 October 2011
Available online 26 November 2011

Keywords:
Remote sensing
Soil moisture
Gap filling
Penalized least square regression
Discrete cosine transform

ABSTRACT

The presence of data gaps is always a concern in geophysical records, creating not only difficulty in interpretation but, more importantly, also a large source of uncertainty in data analysis. Filling the data gaps is a necessity for use in statistical modeling. There are numerous approaches for this purpose. However, particularly challenging are the increasing number of very large spatio-temporal datasets such as those from Earth observations satellites. Here we introduce an efficient three-dimensional method based on discrete cosine transforms, which explicitly utilizes information from both time and space to predict the missing values. To analyze its performance, the method was applied to a global soil moisture product derived from satellite images. We also executed a validation by introducing synthetic gaps. It is shown this method is capable of filling data gaps in the global soil moisture dataset with very high accuracy.

1. Introduction

The presence of data gaps is a cause for concern in many geophysical datasets and presents a large source of uncertainty in data analysis. This is of particular importance when analyzing the spatio-temporal variability of large datasets, e.g., the large-scale satellite observations. In the last two decades satellite observations have demonstrated the potential to become a major tool for observing the properties of the Earth's land surface and atmosphere, such as soil moisture, temperature, aerosols and more recently greenhouse gases. The data gaps in satellite datasets are intrinsic, primarily due to the satellite orbits. Other specific reasons such as clouds contamination or instrumental failure etc can also create data gaps. The rapidly growing volume and diversity of satellite datasets require an efficient method for filling the data gaps.

Several methods for this purpose have emerged in recent years (e.g., Diamantopoulou, 2010), among which the most promising ones are based on the empirical orthogonal function (EOF) of spatial variability (Beckers and Rixen, 2003; Alvera-Azcárate et al., 2007) or the singular spectrum analysis (SSA) of temporal variability (Kondrashov and Ghil, 2006; Hocke and Kämpfer, 2009). These methods use a few spatial or temporal optimal modes occurring at low frequencies to predict the missing values. With the other components discarded as noise, these methods may lead to reduced accuracy of the statistical models fitted to the existing values and consequently the predicted missing values from these models. More importantly, for large spatio-temporal datasets it is of critical importance to utilize information from both spatial and temporal variability to predict the missing values. This demands a method that explicitly takes into account the full three-dimensionality (2-D spatial + time) of the spatio-temporal dataset. However, such a method is still not yet reported to date.

Here we introduce a penalized least square method based on three-dimensional discrete cosine transforms, for the purpose of filling data gaps in large spatio-temporal datasets. To show its performance we apply it to a global soil moisture product derived from satellite images. There are two reasons to choose soil moisture dataset as a primary example. First, soil moisture is one important climate component, which affects the drought and heat conditions of lower atmosphere through partitioning of the available net radiation into latent heat for evaporation and sensible heat for temperature (Koster et al., 2010; Seneviratne et al., 2010). Complete soil moisture datasets are nowadays urgently needed,

* Corresponding author.
E-mail address: g.wang@vu.nl (G. Wang).

1364-8152/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.envsoft.2011.10.015

both for a number of practical applications, such as agriculture and weather forecasting (Varella et al., 2010), as also for increased empirical understanding of the interactions between soil moisture and atmosphere. Secondly, soil moisture exhibits temporally a red spectrum (Wang et al., 2010). This provides a special challenge to the existing gap filling methods utilizing only optimal modes at low frequencies (Kondrashov and Ghil, 2006). It is worth noting that some methods exist that are specifically designed for filling data gaps in high-resolution in-situ soil moisture time series as reviewed in Dumedah and Coulibaly (2011); however, these were not compared to our method, which considers large spatio-temporal satellite products with coarse resolution.

2. Data and method

2.1. Global soil moisture product

We use the VU University-NASA (VUA-NASA) global volumetric soil moisture product ($m^3\,m^{-3}$) derived from the Advanced Microwave Scanning Radiometer-Earth Observing System (Owe et al., 2008). This sensor is mounted on NASA's Aqua satellite and has daily ascending (13:30 equatorial local crossing time) and descending (01:30) overpasses. The surface moisture is retrieved with the Land Parameter Retrieval Model (LPRM) that solves simultaneously for the surface soil moisture and vegetation optical depth (Owe et al., 2008). The LPRM is based on a microwave radiative transfer model for passive microwave images that links surface geophysical variables, i.e. soil moisture, vegetation optimal depth and soil/canopy temperature, to the observed brightness temperatures. The C-band (6.9 GHz) channel is generally used to retrieve soil moisture; and the algorithm switches to X-band (10 GHz) when the C-band is contaminated by Radio frequency interference (RFI) (Njoku et al., 2005). This daily product has been validated extensively over a large variety of land surfaces of sparse to moderate vegetations, showing good agreement with in situ observations (De Jeu et al., 2008; Wagner et al., 2007). It has been shown that the VUA-NASA product outperforms other AMSR-E soil moisture product over various land cover types (Draper et al., 2009; Rüdiger et al., 2009). We apply the gap filling method to both the ascending and descending products for the period 2003–2009, which are gridded at 0.50 degree resolution. Here we show only the results from the ascending product.

2.2. Gap filling method

The method to introduce is a penalized least square regression based on three-dimensional discrete cosine transform (DCT-PLS). The DCT-PLS was originally proposed for automatic smoothing of multidimensional incomplete data (Garcia, 2010a,b), and we adapt it here for the purpose of filling data gaps of spatio-temporal geophysical datasets. The PLS regression is a thin-plate spline smoother for generally one-dimensional data array, which trades off fidelity to the data versus roughness of the mean function. Recently, Garcia (2010a) has demonstrated that the PLS regression can be formulated by the DCT, which expresses the data in terms of a sum of cosine functions oscillating at different frequencies. Since the DCT can be multidimensional, thus the DCT-based PLS regression can be immediately extended to multidimensional datasets. We now give a brief introduction of the DCT-PLS algorithm, and refer the mathematical details to Garcia (2010a).

Let X stand for a spatio-temporal dataset with gaps, and W a binary array of the same size indicating whether or not the values are missing. The DCT-PLS seeks for $\widehat{X}$ that minimizes

$$F\left(\widehat{X}\right) = \left\|W^{1/2}\circ\left(\widehat{X}-X\right)\right\|^2 + s\left\|\nabla^2\widehat{X}\right\|^2, \tag{1}$$

where $\|\cdot\|$ is the Euclidean norm, ∇^2 and $\circ$ stand for the Laplace operator and the Schur (element wise) product, respectively. The s is a positive scalar that controls the degree of smoothing: as s increases, the smoothness of $\widehat{X}$ also increases. The $\widehat{X}$ can be easily achieved by rewriting Eq. (1) with the type II DCT and its inverse (IDCT), which forms

$$\widehat{X} = \mathrm{IDCT}\left(\Gamma\circ\mathrm{DCT}\left(W\circ\left(X-\widehat{X}\right)+\widehat{X}\right)\right). \tag{2}$$

Here, the Γ is a three-dimensional filtering tensor defined by

$$\Gamma_{i_1,i_2,i_3} = \left(1+s\left(\sum_{j=1}^{3}\left(2-\cos\frac{(i_j-1)\pi}{n_j}\right)\right)^2\right)^{-1}, \tag{3}$$

where i_j denotes the ith element along the jth dimension, and n_j denotes the size of X along this dimension.

In Eqs. (2) and (3), the DCT-PLS modeling relies only on the choice of the smoothing parameter s. For the purpose of filling data gaps, this parameter needs to have an infinitesimal value (≈ 0) to reduce the effect of smoothing. A high s value leads to the loss of high frequency components. For a specific s value, the suitability of the derived DCT-PLS model to the existing values can be evaluated by the reconstruction error. We define this as the normalized error between original existing values and their reconstructions:

$$\left\|W^{1/2}\circ\left(\widehat{X}-X\right)\right\| / \left\|W^{1/2}\circ X\right\|. \tag{4}$$

Then the model with defined reconstruction error can then be used to predict the missing values.

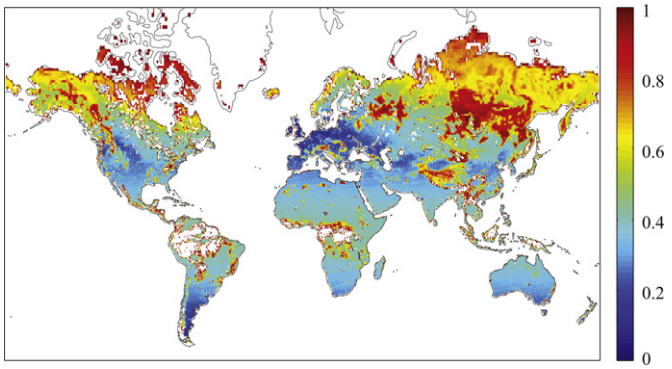

Fig. 1. Fraction of data gaps in the ascending AMSR-E product for the period 2003–2009. White areas contain no data at all.

3. Results

3.1. Gap filling results

Fig. 1 shows the fraction of data gaps that exist in the soil moisture product for the study period. The major reasons that cause data gaps in this dataset include track changes, dense vegetation, frozen soil (snow) and waterbodies. As a polar orbiting satellite, the AQUA satellite gives better coverage over the high latitudes. However, the data gaps amount to 60–90% over north of 45°N because of frozen soil. The same situation also exists for high elevation regions like in the Tibetan Plateau. Over regions of tropical rainforest, the vegetation is too dense to retrieve soil moisture. This product has the best coverage over Europe, with only 10–30% missing values.

Using the DCT-PLS, the approximation of the derived model to the existing values is completely controlled by the smoothing parameter s, which can be any positive value. For the purpose of gap filling rather than smoothing, we consider here only s values much smaller than 1. We apply the DCT-PLS to the global soil moisture product given s values of 10^{-N} with $N = 0, 1, \ldots, 6$ respectively. The global average reconstruction errors for each s value are calculated

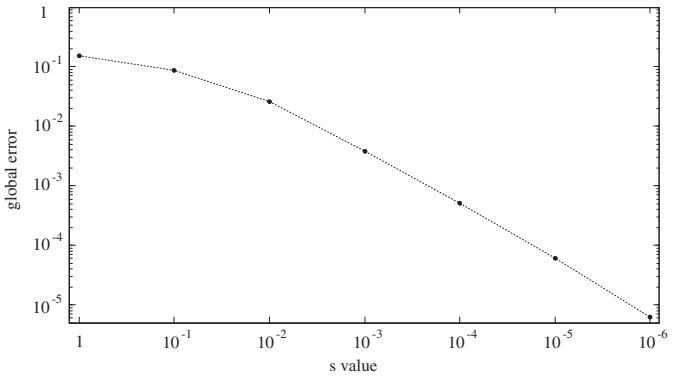

Fig. 2. The reconstruction errors averaged over globe for given s values.

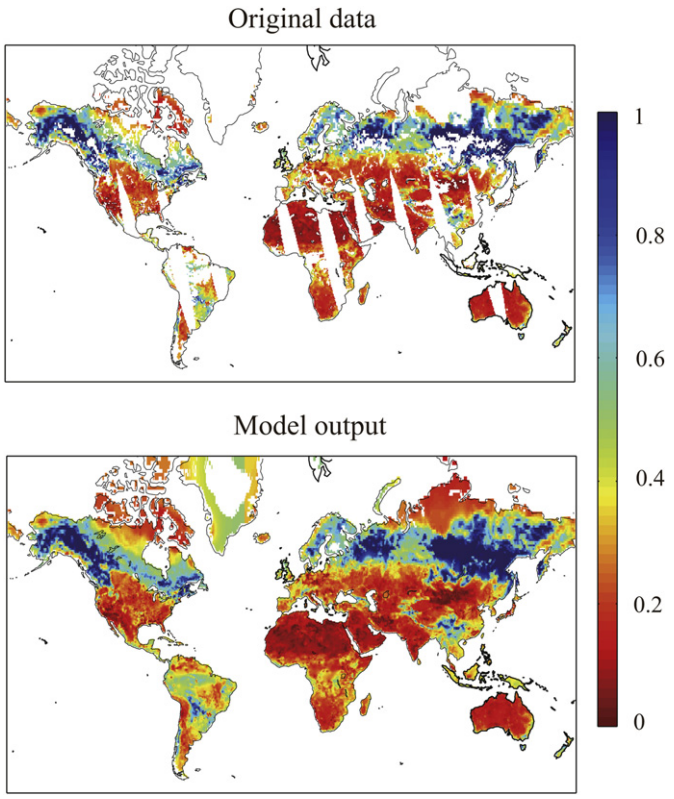

Fig. 3. The data image on Jun. 5, 2003 prior to its model result. Unit: $m^3\,m^{-3}$.

according to Eq. (4), shown in Fig. 2. Not surprisingly, a larger s value results in a larger global error. When $s = 10^{-6}$ is used, the global error has already reached a very small value of 5×10^{-5}. This small error indicates that the derived DCT-PLS model approximates very well the existing values of the global soil moisture dataset; and thus this model can be used to predict the missing values.

Hereafter we demonstrate the gap filling result from the DCT-PLS with $s = 10^{-6}$. We note that the data gaps in the entire dataset are filled by the three-dimensional DCT-PLS simultaneously. The data image and time series shown below are extracted from respectively the original and the gap-filled spatio-temporal datasets. Fig. 3 shows the data image on Jun. 5, 2003 prior to its model result. It appears the missing values are well filled not only between the satellite overpasses but also over the tropical rainforest regions where there are rare observations. In Fig. 4, we show three time series with small to intermediate fraction of data gaps as well as their corresponding model outputs. For a clear presentation, only the data series for 2009 are shown. The upper panel shows the time series extracted from one pixel over Europe (47°N, 2.5°E), with 10% missing values in the original time series. The middle panel shows those from central US (35.5°N, 99°W), with 27% data gaps in the original series. The bottom panel shows those from equatorial Africa (11°N, 0°E), with 43% missing values in the original time series. In all three cases, the original values almost completely overlap their reconstructions by the DCT-PLS model, which is largely due to the small reconstruction error. It is noticeable that the extreme values existing in the original dataset are also well captured by the model; those are emphasized with arrows in Fig. 4. This indicates that the predicted missing values from the used DCT-PLS model indeed might be reliable; however, further validation is shown in section 3.2.

With conventional methods, the hardest part is to fill the continuous gaps. In spatio-temporal dataset the spatially continuous gaps can be temporally intermittent, or vice versa, such as those between the satellite overpasses. Owing to the three-dimensionality, the DCT-PLS method can easily cope with data gaps of this type. However, special attention needs to be paid to data gaps of large spatio-temporal size, e.g., those over the tropical rainforest regions where the vegetation is too dense to retrieve soil moisture. In this case, the missing values are predicted using the low frequency components of the dataset. This leads to reduced reliability of filled-in high frequency components. A large portion of data gaps of this global soil moisture dataset is due to frozen soil, in which case the filled-in soil moisture values are physically not realistic.

3.2. Synthetic validation

Sometimes perfect fitting does not necessarily imply good prediction skill, for example, when overfitting occurs. Thus the

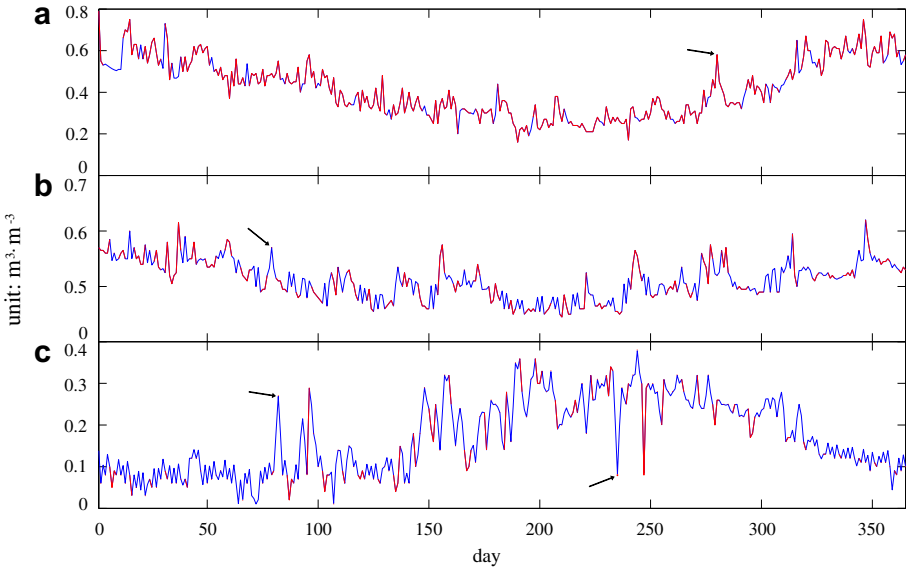

Fig. 4. Original values (red) prior to their corresponding model reconstructions (blue) for the year 2009 from the pixels over a. Europe (47°N, 2.5°E), b. US (35.5°N, 99°W) and c. Africa (11°N, 0°E). Note that the original values are almost completely overlapped by the reconstructed values, due to the very small reconstruction errors. Emphasized with arrows are some extreme values.

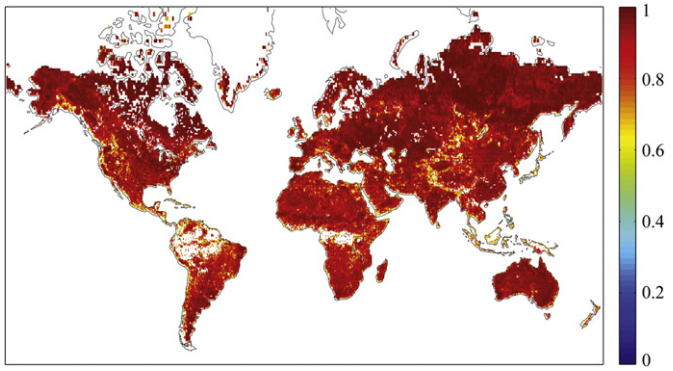

Fig. 5. Pixel-wise *Corr* ($p < 0.05$) surface for the synthetic validation over globe.

prediction skill needs to be further validated, for which we use a generally accepted approach. To validate the prediction skill of the DCT-PLS method, we introduce synthetic gaps in addition into the original soil moisture dataset (2003–2009) by randomly removing 10% of the existing values pixel by pixel. This validation strategy is oriented to the data gaps which are temporally intermittent. In case of this data intermittency, creating synthetic gaps at random is the most efficient and realistic as we can not assume any particular distributions of the data gaps. Then the DCT-PLS gap filling process is applied to the new dataset with $s = 10^{-6}$. In the synthetic gaps, we calculate the correlation coefficient (*Corr*) between the original values and their corresponding DCT-PLS predictions. This is, the error statistics are only calculated for the data gaps that were synthetically created. The reconstruction error in the synthetic gaps can be alternatively used as the skill metric of prediction; however, it contains no more information than *Corr*, and we show here only *Corr*. The pixel-wise *Corr* ($p < 0.05$) is shown in Fig. 5. It appears that 85% of the validated pixels have higher *Corr* than 0.80, and those pixels with higher *Corr* than 0.90 amount to 64%. Specifically, the *Corr* values for the representative cases in Fig. 4 are 0.97 (Europe), 0.95 (US) and 0.97 (Africa) respectively. This suggests very good prediction skill of the DCT-PLS for the filling the data gaps of spatio-temporal soil moisture dataset.

4. Discussion

In this communication, we introduce an efficient DCT-PLS method for filling the data gaps in large spatio-temporal dataset. It is recommended particularly for the rapid growing volume and diversity of satellite observations in environmental sciences. Using a global satellite soil moisture dataset as example and as challenging case, we have demonstrated the very good skill of this method for gap filling purposes.

This DCT-PLS method has some novel features with respect to other gap filling methods. It is a method of full three-dimensionality, and thus explicitly utilizes both spatial and temporal information of the dataset to derive the statistical model and predict the missing values. Instinctively, this strategy is preferable for spatio-temporal datasets rather than using only spatial or temporal modeling. The statistical modeling process is completely controlled by one smoothing parameter which is easy to specify and eliminates the need for complicated model parameterizations. Furthermore, with a small smoothing parameter the DCT-PLS method has the potential to reliably fill in the high frequency components.

However, there is no fixed relation between the smoothing parameter and the gap filling result. In the case where the geophysical datasets have spatially very large differences in magnitude, an overfitting might occur with an extremely small smoothing parameter, leading to poor prediction performance. For example, for our soil moisture data the minor fluctuations in the dataset are indeed observed to be exaggerated over some regions, when a smoothing parameter smaller than 10^{-7} is used. There are alternative ways to avoid the overfitting problem whether or not the dataset contains continuous spatio-temporal gaps of large size. For datasets without such gaps, the best choice is probably to introduce cross validation for better generalization as Garcia (2010a) suggested. Yet, this may lead in turn to underfiting and errorous prediction where data gaps of large spatio-temporal size exist. In this case we suggest a post-validation by introducing synthetic gaps to ensure the reliability of the filled in values. The Matlab code for this method is available from: http://www.biomecardio.com/matlab/smoothn.html.

Acknowledgements

The work was supported by the The Netherlands Organization for Scientific Research (NWO; grant No. 854.00.026) and ESAs STSE funded Integrated Project WAter Cycle Multimission Observation Strategy (WACMOS; Contract No. 22086/08/I-EC).

References

Alvera-Azcárate, A., Barth, A., Beckers, J.M., Weisberg, R.H., 2007. Multivariate reconstruction of missing data in sea surface temperature, chlorophyll, and wind satellite fields. J. Geophys. Res. 112, C03008. doi:10.1029/2006JC003660.
Beckers, J., Rixen, M., 2003. EOF calculations and data filling from incomplete oceanographic data sets. J. Atmos. Ocean. Technol. 20, 1839–1856.
De Jeu, R.A.M., Wagner, W.W., Holmes, T.R.H., Dolman, A.J., van de Giesen, NC, Friesen, J., 2008. Global soil moisture patterns observed by space borne microwave radiometers and scatterometers. Surv. Geophys. 29, 399–420.
Diamantopoulou, M.J., 2010. Filling gaps in diameter measurements on standing tree boles in the urban forest of Thessaloniki, Greece. Environ. Model. Softw. 25, 1857–1865. doi:10.1016/j.envsoft.2010.04.020.
Draper, C.S., Walker, J.P., Steinle, P.J., de Jeu, R.A.M., Holmes, T.R.H., 2009. An evaluation of AMSR-E derived soil moisture over Australia. Remote Sens. Environ. 113 (4), 703–710.
Dumedah, G., Coulibaly, P., 2011. Evaluation of statistical methods for infilling missing values in high-resolution soil moisture data. J. Hydrol. 400, 95–102.
Garcia, D., 2010a. Robust smoothing of gridded data in one and higher dimensions with missing values. Comput. Stat. Data. Anal. 54, 1167–1178.
Garcia, D., 2010b. A fast all-in-one method for automated post-processing of PIV data. Exp. Fluids. doi:10.1007/s00348-010-0985-y.
Hocke, K., Kämpfer, N., 2009. Gap filling and noise reduction of unevenly sampled data by means of the Lomb-Scargle periodogram. Atmos. Chem. Phys. 9, 4197–4206. doi:10.5194/acp-9-4197-2009.
Kondrashov, D., Ghil, M., 2006. Spatio-temporal filling of missing data points in geophysical data sets. Nonlin. Process. Geophys. 13, 151–159.
Koster, R.D., et al., 2010. Contribution of land surface initialization to subseasonal forecast skill: first results from a multi-model experiment. Geophys. Res. Lett. 37, L02402. doi:10.1029/2009GL041677.
Njoku, E., Ashcroft, P., Chan, T.K., Li, L., 2005. Statistics and global survey of radio-frequency interference in AMSR-E land observations. IEEE Trans. Geosci. Remote Sens. 43, 938–947.
Owe, M., de Jeu, R., Holmes, T., 2008. Multisensor historical climatology of satellite-derived global land surface moisture. J. Geophys. Res. 113, F01002. doi:10.1029/2007JF000769.
Rüdiger, C., Calvet, J.C., Gruhier, C., Holmes, T., de Jeu, R., Wagner, W., 2009. An intercomparison of ERS-Scat and AMSR-E observations, and soil moisture simulations over France. J. Hydrometeorol. 10 (2), 431–447.
Seneviratne, S.I., Corti, T., Davin, E.L., Hirschi, M., Lehner, I., Teuling, A.J., 2010. Investigating soil moisture–climate interactions in a changing climate: a review. Earth Sci. Rev. 99, 125–161.
Varella, H., Guérif, M., Buis, S., 2010. Global sensitivity analysis measures the quality of parameter estimation: the case of soil parameters and a crop model. Environ. Model. Softw. 25, 310–319.
Wagner, W., Naeimi, V., Scipal, K., de Jeu, R., Martinez-Fernandez, J., 2007. Soil moisture from operational meteorological satellites. Hydrogeol. J. 15 (1), 121–131.
Wang, G., Dolman, A.J., Blender, R., Fraedrich, K., 2010. Fluctuation regimes of soil moisture in ERA40 re-analysis dataset. Theor. Appl. Climatol. 99, 1–8.