Spatiotemporal reconstruction of gaps in multivariate fields using the direct sampling approach

Gregoire Mariethoz,[1] Matthew F. McCabe,[1] and Philippe Renard[2]

Received 12 March 2012; revised 7 July 2012; accepted 30 August 2012; published XX Month 2012.

[1] The development of spatially continuous fields from sparse observing networks is an outstanding problem in the environmental and Earth sciences. Here we explore an approach to produce spatially continuous fields from discontinuous data that focuses on reconstructing gaps routinely present in satellite-based Earth observations. To assess the utility of the approach, we use synthetic imagery derived from a regional climate model of southeastern Australia. Orbital tracks, scan geometry influences, and atmospheric artifacts are artificially imposed upon these model simulations to examine the techniques' capacity to reproduce realistic and representative retrievals. The imposed discontinuities are reconstructed using a direct sampling technique and are compared against the original continuous model data: a synthetic simulation experiment. Results indicate that the multipoint geostatistical gap-filling approach produces texturally realistic spatially continuous fields from otherwise discontinuous data sets. Reconstruction results are assessed through comparison of spatial distributions, as well as through visual assessment of fine-scale features. Complex spatial patterns and fine-scale structure can be resolved within the reconstructions, illustrating that the often nonlinear dependencies between variables can be maintained. The stochastic nature of the methodology makes it possible to expand the approach within a Monte Carlo framework in order to estimate the uncertainty related to subsequent reconstructions. From a practical perspective, the reconstruction method is straightforward and requires minimum user intervention for parameter adjustment. As such, it can be automated to systematically process real time remote sensing measurements.

Citation: Mariethoz, G., M. F. McCabe, and P. Renard (2012), Spatiotemporal reconstruction of gaps in multivariate fields using the direct sampling approach, *Water Resour. Res.*, *48*, XXXXXX, doi:10.1029/2012WR012115.

1. Introduction

[2] Earth observation, whether from in situ networks, intensive (but usually sporadic) field campaigns, or from satellite-based remote sensing retrievals, provides an inherently discontinuous stream of data. Remote sensing based retrievals of the terrestrial, ocean, atmosphere and subsurface systems, have significant potential to inform a variety of Earth system modeling applications [*Brunner et al.*, 2004; *Li et al.*, 2009; *Milzow et al.*, 2009]. From a terrestrial hydrological perspective, satellite retrievals have been used to characterize spatially and temporally varying fields such as soil moisture [*Jeu et al.*, 2008; *Liu et al.*, 2011], evapotranspiration [*Kalma et al.*, 2008; *Su et al.*, 2007], rainfall [*Huffman et al.*, 1995; *Kummerow et al.*, 2000], radiation [*Diak et al.*, 1996; *Weymouth and Le Marshall*, 2001] and even seek observationally based hydrological closure [*Sahoo et al.*, 2011; *Sheffield et al.*, 2009]. However, one of the confounding problems with the use of such observations is the presence of spatial discontinuities, due to incomplete coverage of the domain resulting from satellite orbital characteristics or through occlusion by cloud cover and other atmospheric effects. Such discontinuities often make satellite-based observations difficult to integrate within traditional modeling frameworks, which prefer spatially and temporally continuous data fields.

[3] The problem of gap filling in spatially discontinuous data sets, including those inherent in retrievals from Earth observing systems, has been the focus of many research investigations [e.g., *Boloorani et al.*, 2008; *Maxwell et al.*, 2007; *Wang et al.*, 2012; *Yuan et al.*, 2011; *Zhang et al.*, 2007]. In general terms, the gap-filling problem can be formulated as determining the value of a pixel with spatial constraints (it must be coherent with the surrounding values), temporal constraints (it must be coherent with the preceding values), and also constraints related to any dependence with covariates (which may not necessarily be linear dependencies). For example, topography is a covariate which is known to be influential on the spatial distribution of rainfall and air temperature [*Goovaerts*, 2000]. In many gap-filling studies, the reconstruction problem is often relatively well defined due to one or more of the following reasons: (1) the

[1]School of Civil and Environmental Engineering, University of New South Wales, Sydney, New South Wales, Australia.
[2]Centre of Hydrogeology and Geothermics, University of Neuchatel, Neuchatel, Switzerland.

Corresponding author: G. Mariethoz, School of Civil and Environmental Engineering, University of New South Wales, Sydney, NSW 2052, Australia. (gregoire.mariethoz@unsw.edu.au)

0043-1397/12/2012WR012115

variable of interest is available at a time step that is close, relative to the temporal variability of the studied phenomenon, (2) the spatial extent of the gaps is small relative to the size of the features being reconstructed, (3) some strongly informative or linearly correlated covariates are available, and (4) there is only a single unknown variable to reconstruct, therefore the problem of preserving relationships between different uninformed variables does not exist. These types of problems can be described as strongly constrained gap filling, because the amount of information available to guide the interpolation is considerable. In such cases, some relatively simple methods such as image compositing can successfully address the problem [*Cihlar*, 2000; *Du et al.*, 2001]. While cokriging generally gives better results than image compositing, the highly constrained nature of the problem is similar [*Pringle et al.*, 2009].

[4] In this paper, we address a more challenging class of problems, referred to herein as weakly constrained gap-filling problems. Weakly constrained problems would include phenomena that change at subdaily time scales, where exhaustive measurements during preceding days may not be available or informative enough to fill gaps on some other day, and where filling cannot be inferred from linear correlation with a covariate. Another characteristic of weakly constrained gap filling is the significant size of the gaps compared to the size of the structures present in the image.

[5] As a result of the weak constraints imposed on the interpolation problem, the solution is necessarily nonunique, stressing the need to quantify uncertainty. In previous studies, the high determinism of so-called strongly constrained gap-filling problems did not generally confront the question of uncertainty in the interpolation results. We adopt the framework of geostatistics, which offers the means of evaluating interpolation uncertainty, either through an estimation variance in the case of kriging, or through Monte Carlo analysis.

[6] A popular approach to gap filling is cokriging [*Zhang et al.*, 2012]. *Zhang et al.* [2009] applied the technique to multispectral images to impose correlation with the same gap free image taken four months earlier. However, kriging and its variations have two major limitations: they are smoothing interpolators and can only account for linear relations with covariates [*Chilès and Delfiner*, 1999; *Goovaerts*, 1997]. Using kriging can result in the interpolated areas being visibly distinct from the rest of the image, presenting unrealistic continuous textures and, if point measurements are available, artifacts near these locations. Other geostatistical techniques such as stochastic simulation allow for better representing the textures present in the data. However, the underlying models are based on two-point statistics and may therefore not reproduce the complex spatial patterns present in known parts of the domain [*Journel and Zhang*, 2006]. Moreover, when dealing with multiple variables, these methods often consider linear relationships, which are oversimplifications in many environmental modeling applications [*Rivoirard*, 2001].

[7] In this paper we investigate the use of multiple-point geostatistics for gap-filling applications. The method employed here is the direct sampling approach [*Mariethoz and Renard*, 2010; *Mariethoz et al.*, 2010]. The technique exploits intrinsic relationships between linked observations and offers the capacity to provide more realistic spatially continuous fields from remote sensing based platforms and broaden their effective use and integration within the Earth sciences. Multiple-point geostatistics methods use training images to describe a time varying data set for periods other than the missing time, which then allows for the identification of specific spatial patterns that might be expected to recur in subsequent scenes. The spatial patterning and image structure can then be used to improve the gap-filling procedure. The supplementary use of multiple covariates, which are themselves incomplete, is at the foundation of the approach presented in this paper.

[8] In this preliminary assessment of the technique to Earth system data, we apply the reconstruction method to regional climate model (RCM) simulations over southeastern Australia [*Evans and McCabe*, 2010] and use this as a synthetic surrogate for remote sensing based retrievals. The advantage of using synthetic model output as opposed to actual satellite data is the capacity (1) to artificially impose distributions of gaps that can reflect both expected orbital features and atmospheric condition and (2) to assess subsequent image reconstructions against a spatially continuous modeled "truth." An especially important aspect of using synthetic imagery is that it ensures we address a weakly constrained problem, by imposing gaps typically larger than the spatial structures present. It also allows the production of data sets where gaps occult simultaneously across multiple nonlinearly related variables, and to then validate the results against the known nonlinear relationship. Such a validation would be extremely difficult with real data.

[9] In the sections 2 and 3 we detail the structure and logic behind the direct sampling approach and then develop realistic scenarios based on these synthetic data to describe and assess the potential application of the technique to remote sensing retrievals.

2. Methodology

[10] The direct sampling algorithm [*Mariethoz et al.*, 2010] generates stochastic fields that can present complex statistical and spatial properties. These properties are usually inferred from a fully informed training image, but it has been shown that the method can also be used without training images if a large portion of the domain is already known. In the latter case, the incomplete image is reconstructed by using patterns borrowed from the known parts of the image. In the past, the method has been successfully applied to reconstruct 3-D geological structures and borehole images [*Mariethoz and Renard*, 2010]. In this paper we propose to use a conceptually similar approach to reconstruct missing regions of multivariate synthetic satellite data, based on values that are known at different locations and/or different dates. Although we focus on the multivariate aspect, it should be noted that the entire methodology is equally applicable to univariate images, which is a simpler case.

[11] As an example of the approach, consider a thermal infrared image affected by missing data resulting from cloud cover or other atmospheric contamination. A temporally coincident retrieval unaffected by clouds (e.g., derived from a microwave sensor) is likely to be available from a variety of other satellites, or even from additional sensors present on the same platform [*McCabe et al.*, 2008a]. The

information need not even be the same variable, but may just possess some statistical or physical relationship with the missing variable of interest. For instance, there will be locations where soil moisture retrievals will not be informed due to a scan gap in the orbiting sensor, but where another satellite might provide information on the surface temperature, radiation, or even both of these observations. The additional information that is available to inform upon the missing value, may then be related to the spatial structure (or texture) of that variable. For instance, it might be inferred from past observations that the surface temperature presents large variations at relatively small spatial scales, whereas soil moisture values are spatially more continuous. This level of information can be taken into account to determine the nature of variability to reproduce, and the relative influence of data around the gap locations. The approach proposed here seeks to use this textural relationship and the dependency to coincident and noncoincident complementary information, by exploiting the different textural properties and their interrelationships.

[12] To accurately convey the concept of an event that is situated in both spatial and temporal terms, we use the terminology location/date. Let $\mathbf{x}$ be a pixel in the image where the variable of interest $Z(\mathbf{x})$ is uninformed and needs to be reconstructed. We denote $\mathbf{N}_x$ as the ensemble of the n closest pixels of $\mathbf{x}$ that are informed. The basic idea of the reconstruction algorithm is to find another location/date $\mathbf{y}$ in the image that is informed and that has a set of values in the neighborhood $\mathbf{N}_y$ similar to those in the neighborhood $\mathbf{N}_x$. When a suitable location/date $\mathbf{y}$ is found, its value $Z(\mathbf{y})$ is assigned to $Z(\mathbf{x})$. The main conceptual caveat in this procedure is that we are not interested in the location/date of maximum similarity for two major reasons: (1) it would result in a deterministic value for $Z(\mathbf{x})$, therefore not allowing one to quantify the uncertainty of the reconstructed values, and (2) it would involve a history search on the entire image for each location/date $\mathbf{x}$ to be reconstructed, thus involving a large computational load.

[13] Instead, we want one possible outcome of Z conditioned to $\mathbf{N}_x$, i.e., a sample of the conditional cumulative distribution function:

$$F(z) = \text{Prob}(Z(\mathbf{x}) \leq z | \mathbf{N}_x). \quad (1)$$

[14] The direct sampling approach accomplishes this by performing a Shannon-type sampling [*Shannon*, 1948] by scanning the training data set and computing, at each location, the distance $d(\mathbf{N}_x, \mathbf{N}_y)$. It is based on the principle that the *1st* random location/date $\mathbf{y}$ encountered in the training data set, whose neighborhood is sufficiently similar to the one of $\mathbf{x}$, is necessarily a sample of $F(z)$. The similarity between neighborhoods $\mathbf{N}_x$ and $\mathbf{N}_y$ relies on the use of the distance $d(\mathbf{N}_x, \mathbf{N}_y)$. A brief discussion of the most important aspects of this method is given below. A detailed description of the algorithm is provided by *Mariethoz et al.* [2010] for further reference.

[15] Since the distances are usually defined such that they are within the interval [0,1], the threshold t is also bound to the same interval. Defining a threshold of $t = 0$ means that the patterns of the training image will be reproduced with the highest possible accuracy, and the method is then essentially data driven. Conversely, when setting $t = 1$ the algorithm unconditionally samples values from the training image, and therefore only reproduces the marginal distribution of Z without any constraints in terms of spatial dependence. Between these two extreme cases, the value of t determines how accurately the patterns of the training image are reproduced. In general, increasing t relaxes the constraints on the spatial dependence of the reconstructed fields and eases the computational burden. The parameter is usually adjusted by performing a sensitivity study on a small-scale model.

[16] The distance $d(\mathbf{N}_x, \mathbf{N}_y)$ can be computed in different ways, depending on the nature of the variable to reconstruct (for a discussion of the different possible distances to use, see *Mariethoz et al.* [2010] and *Mariethoz and Kelly* [2011]). Distances have been proposed to be used with both categorical and continuous variables. In this paper, we consider synthetic satellite images consisting of continuous variables and therefore adopt the pair wise Manhattan distance to quantify the dissimilarity between the values of any two neighborhoods:

$$d(\mathbf{N}_x, \mathbf{N}_y) = \frac{1}{\eta} \langle |\mathbf{N}_x - \mathbf{N}_y| \rangle. \quad (2)$$

where $\langle\,\rangle$ denotes the average and η is a normalization factor ensuring that the distance values remain bounded between 0 and 1. A usual value for η is the maximum difference between two values of Z observed in the training image or training data. While Manhattan and Euclidean distances both yield very similar results, the Manhattan distance was chosen here because it is slightly less computationally demanding.

[17] An important point to consider is that several unique satellites systems may provide different pieces of information for a particular study area. For example, there may be locations where, at a given date, hydrological variables such as the rainfall and the soil moisture have been recorded, but the evaporative flux has not. At other locations, depending upon the areas covered by each satellite, there may be any number of variables informed by other independent data sets. If several variables inform upon areas requiring reconstruction, the direct sampling method defines neighborhoods spanning across the different variables, and uses a specific distance measure. One then needs to consider separately the neighborhood of $\mathbf{x}$ for each of the m variables Z^k, $k = 1 \ldots m$. Then, $\mathbf{N}'_x$ is the multivariate neighborhood of $\mathbf{x}$, constituted by all subneighborhoods $\mathbf{N}^k_x$ taken together:

$$\mathbf{N}'_x, \mathbf{N}^1_x \cup \ldots \cup \mathbf{N}^k_x \cup \ldots \cup \mathbf{N}^m_x. \quad (3)$$

[18] The distance used to compute similarity between multivariate neighborhoods is a weighted average of the distances taken individually for each univariate neighborhood:

$$d'(\mathbf{N}'_x, \mathbf{N}'_y) = \sum_{k=1}^{m} \frac{w_k}{\eta_k} \langle |\mathbf{N}^k_x - \mathbf{N}^k_y| \rangle, \; \sum_{k=1}^{m} w_k = 1, \quad (4)$$

where w_k are the weights and η_k the normalization constants for each variable. Figure 1 presents a graphical

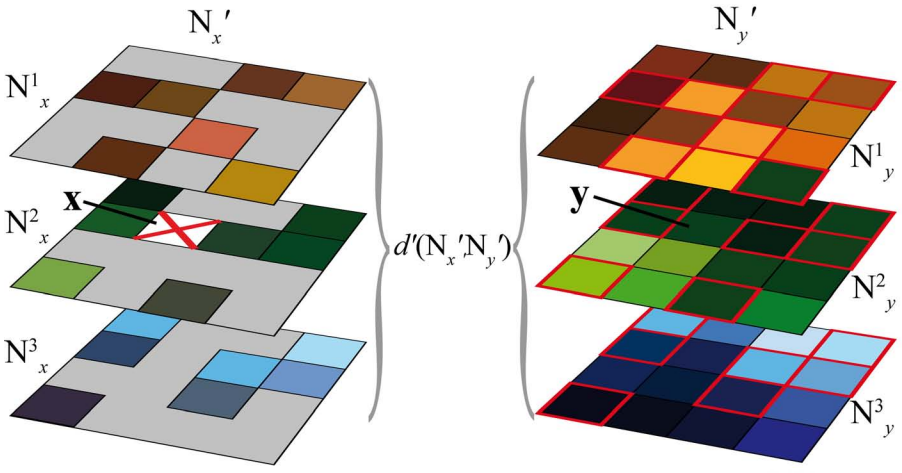

Figure 1. Multivariate neighborhoods and multivariate distances for a case with three variables. (left) The uninformed location/date x to reconstruct is contained within the second variable. Informed locations/dates are colored, and uninformed locations/dates are in gray. (right) An informed part of the domain. The distance between both is computed using the locations/dates marked in red, which correspond to the informed locations in the neighborhood of x.

representation of the principle of multivariate neighborhoods and illustrates the computation of the distance between these.

[19] The two most important parameters of the direct sampling algorithm are the size of the neighborhoods considered and the threshold value t. Tests have showed that larger neighborhoods (at least 20 pixels) allow for generation of the complex patterns and shapes found in natural images. However, using too large a search neighborhood may lead to a dramatic increase in computation cost. Typical neighborhoods are constituted of between 20 and 40 pixels.

3. Application to Synthetic Data

[20] The reconstruction method is applied to data derived from a regional climate model of southeastern Australia. Simulations from the Weather Research and Forecasting (WRF) model (see *Evans and McCabe* [2010] for a complete model description) were used as a proxy for satellite-based retrievals of common land surface variables. WRF is a widely utilized coupled numerical model used to describe land-atmosphere interactions. However, its use here is relevant only in that it produces spatially continuous fields that represent physically consistent descriptions of water and energy cycle behavior in a hydrologically consistent manner [*McCabe et al.*, 2008a]. An initial focus period of 365 days was identified from the longer-term simulations, with individual daytime (12 PM) reproductions extracted from the model, representing a spatial resolution of 10 km and a domain size of 243 $\times$ 186 pixels. Spatial fields of soil moisture, surface temperature, latent heat flux and shortwave downward radiation (a surrogate for cloud coverage) were extracted from the WRF output for further analysis.

[21] Apart from geostationary satellites, which sense whole Earth disk images at regular time intervals, most commonly used Earth Observing Systems (EOS) form part of a constellation of polar orbiting platforms. An inherent feature of these systems is the progressive development of scan tracks in response to the orbital geometry of the satellite and sensor characteristics, which routinely result in regions on the curved Earth surface that are not measured during the satellite overpass. For optical and infrared-based sensors, atmospheric effects manifested by cloud cover and other meteorological phenomena present a recurring problem, particularly when the orbital characteristics limit coverage to subdaily overpasses (i.e., most polar orbiting sensors). Developing a semiphysical basis from which missing data might be reconstructed would provide considerable advantage for extending the utility of satellite-based retrievals.

[22] Two scenarios are considered. In the first, artificial gaps corresponding to exaggerated satellites scan tracks are imposed upon the images, masking an area of up to 40% of the domain. These artificial gaps are then reconstructed with direct sampling and compared with the original model output. It is assumed that each variable has different gaps, as in practice they might be informed by a different satellite. In the second scenario, gaps are positioned to correspond with the highest density of clouds (as simulated by the WRF model) to reflect the influence of atmospheric features on the satellite observation. For both scenarios, gaps occur in all variables except for the downward shortwave radiation, which is known everywhere in the image: a reasonable assumption based on its retrieval from geostationary platforms.

[23] Climate variables generally present temporal nonstationarity, i.e., the statistics and patterns of values found in summer and winter differ and should be treated separately. For this reason, we reconstruct the values separately for each month, using only the known values in January to reconstruct the gaps in January, and likewise for other months. In this case the informed parts of the domain are large enough to ensure proper temporal correlation across the different months. If it was not the case, one could devise a moving window scheme whereby the values of the 15 days before and after the present date are used as training data set. For the direct sampling, we use neighborhoods consisting of the $n = 20$ locations/dates that are the closest to $\mathbf{x}$, for each of the 4 variables considered, thus producing

a total of 80 pixels. The distance function in equation (4) is used with a threshold of 0.01. Each of the 4 variables is given an equal weight in the distance calculation.

4. Results

4.1. Imposing Satellite Scan Tracks on Synthetic Data

[24] One of the most predictable outcomes of Earth observing systems is the development of scan tracks as the satellite overpass progresses through its orbit. These discontinuities, a result of compromises between a satellites field of view, repeat rate, and orbital characteristics, are more pronounced along the equatorial region: at least for polar orbiting systems, where the spatial separation between orbits is often the greatest. In order to reproduce the effect of these gaps for subsequent reconstruction, an exaggerated response is imposed upon the synthetic fields, with diagonal scan lines enforced upon each of the variables of interest. While the gaps are wider and cover significantly more area than would actually occur in practice, a key motivation of this paper is to rigorously examine the utility of the direct sampling approach to remote sensing data reconstruction. The use of such large gaps allows for testing of the method under adverse conditions, exploring the limits of its application.

[25] As noted previously, the direct sampling technique is not a deterministic approach, but rather allows for the assessment of inherent uncertainties in reconstructing possible spatial fields. To address this, 10 stochastic reconstructions (or realizations) have been computed, providing 10 possible values at each unknown pixel for all the individual WRF model simulations. In this case, using a larger number of realizations was compromised against the computational constraints due to the high temporal resolution of model reproductions. Each time step is reconstructed based on an entire month of data, resulting in a large training set. For simplicity we display detailed results for single representative images taken at different periods in the year, along with monthly statistics.

[26] Figure 2 presents a sample realization for 15 January 2006. For the three variables that are partially informed, describing surface latent heat (LH), surface temperature (TSK), and soil moisture (SMOIS), we display the incomplete data, a direct sampling realization, and the reference truth. Shortwave downward radiation (SWDOWN) is considered spatially continuous (i.e., it has no gaps) and is shown separately.

[27] The patterns in the reconstructed variables seem realistic, especially given the significant amount of gaps present in the images. The large-scale structures are maintained and qualitative agreement is satisfactory: although discontinuities appear for surface temperature reconstructions. Figure 3 plots the statistics of all reconstructions for the entire month of January versus the reference WRF data sets. Figure 3 (top) shows that the direct sampling reconstructions are well correlated with the real values. In Figure 3 (bottom), the histograms of errors are displayed (i.e., the reconstructed values minus the real ones). These errors are centered on zero and mostly unbiased, apart from the soil moisture results, where anomalies are caused by the highly irregular distribution of the values. For latent heat flux and soil moisture, the errors fall within expected ranges for satellite retrievals. In terms of root mean square error, these have been estimated previously as anywhere between 20–100 W m^{-2} for LH [*Kalma et al.*, 2008; *Kustas and Norman*, 2000] and 3–7% for SMOIS [*Drusch et al.*, 2004; *Liu et al.*, 2011; *McCabe et al.*, 2005] for instantaneous retrievals. For surface temperature (TSK), errors are broadly similar to typical measurement errors of between 1 and 8 K [*Ferguson and Wood*, 2010; *McCabe et al.*, 2008b; *Wan et al.*, 2002]. These results compare favorably to the gap-filling errors shown in [*Chen et al.*, 2011], especially when one considers that in our case, three interdependent variables are dealt with simultaneously instead of just one. The spatial distribution of interpolation errors are more clearly observed in Figure 4, which displays the ensemble statistics of all 10 realizations, with the mean value and standard deviation at each reconstructed pixel, along with the average absolute error.

[28] Figures 5–7 show similar results for a midwinter analysis. Figures 5 and 7 correspond to the reconstruction results for 15 July, while Figure 6 shows reconstruction errors for the entire month of July. A significant difference with the results of January is a smoother error distribution for soil moisture. This can be explained by the presence of persistent features along drainage systems during the summer (January), which are larger than the size of gaps. This involves nonstationarity within gaps, which can lead to decreased pattern reproduction. In the wetter winter months (July), the soil moisture is more affected by rainfall and hence the patterns tend to have shorter correlation scales, making the stationarity assumption more likely to be valid.

[29] The four variables (LH, TSK, SMOIS, and SWDOWN) present nonlinear relationships with each other, as shown in the scatterplots in Figure 8 (first and second rows) that corresponds to the entire month of January. Note that while the data could have been displayed as joint distributions, scatterplots are used because the large amount of data allows for good visual representation. Only the pixels simulated at the gap locations have been considered in Figure 8. Reproducing such joint distribution is typically difficult to achieve and only a few methods currently allow approximating them [*Leuangthong and Deutsch*, 2003; *Mariethoz et al.*, 2009; *Yeh et al.*, 1996]. Figure 8 (third and fourth rows) displays the same series of scatterplots, but using locations from the reconstructed images. The reconstructed distributions are in excellent agreement with the reference, providing confidence that the complex dependencies observed in the reference can be satisfactorily reproduced: a consequence of the direct sampling method considering entire multivariate patterns.

[30] However, the price to pay for this accurate multivariate dependency may come at the cost of a loss in the spatial continuity of the reconstructed variables. Indeed, some reconstructions can present artifacts at the boundaries of the reconstructed domain for surface temperature (Figure 5). These discrepancies can occur when relatively small distances are obtained because of a good match to some of the variables, despite having a poor match to the actual variable being simulated. In other words, with the distance in equation (4), it may be possible to find a neighborhood $\mathbf{N}_y$ having a sufficiently small distance with $\mathbf{N}_x$, but where all the similarity is due to three of the four variables (e.g., LH, SMOIS, SWDOWN), with the fourth variable (e.g., TSK) presenting significant discrepancies. In such cases, the spatial correlation

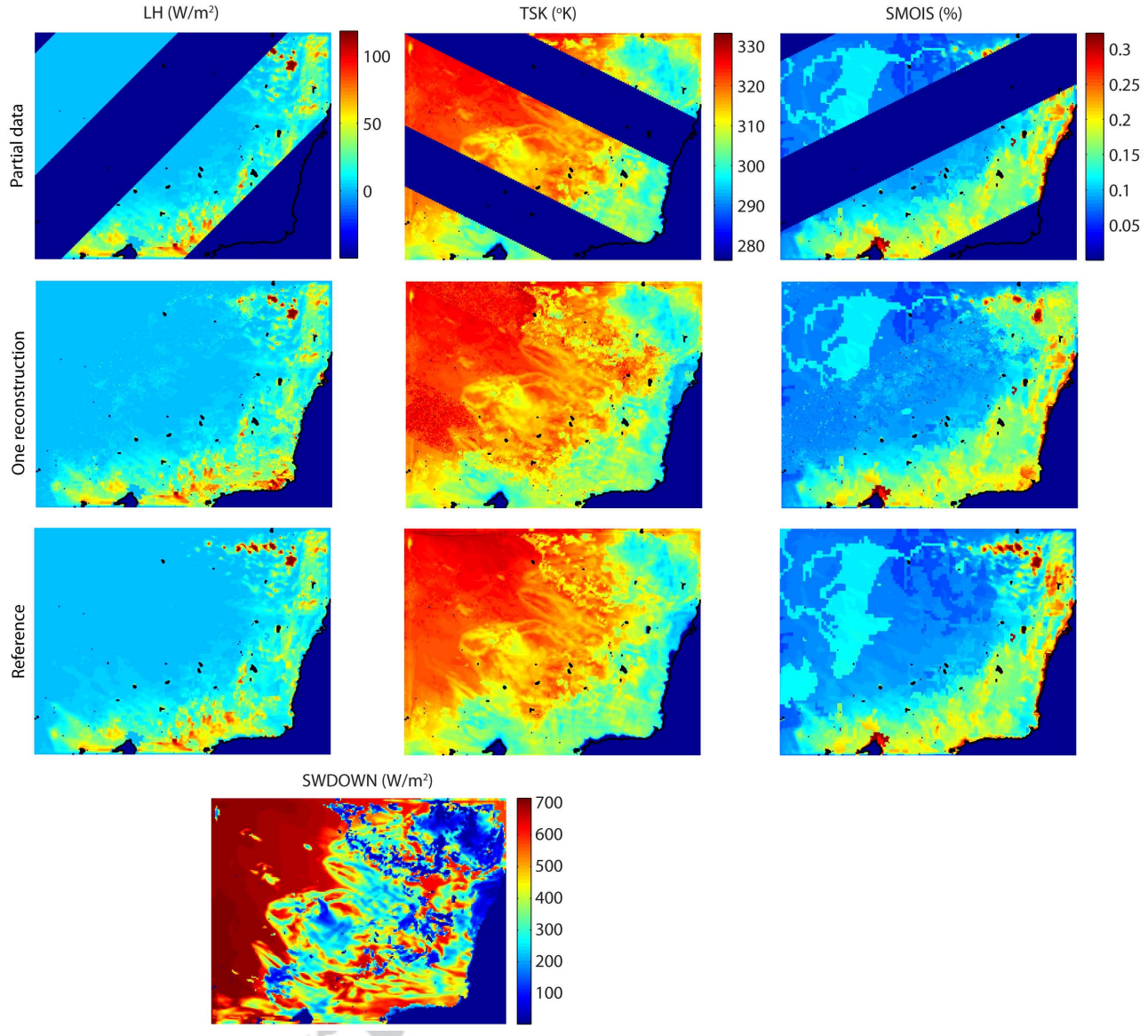

Figure 2. Gaps caused by orbital characteristics in a sample reconstruction for 15 January 2006. Three columns representing WRF simulated data fields are (from left to right) latent heat flux (LH), surface temperature (TSK), and soil moisture (SMOIS). The rows detail (from top to bottom) the artificially gap enforced simulation, the reconstructed image, and the original continuous WRF simulation. Downward shortwave radiation is included at the bottom, with the influence of cloud evident throughout the bottom left portion of the image.

of the fourth variable (TSK in this example) has insufficient weight in the computation of the distance, leading to the observed spatial discontinuities. These issues, and potential ways to resolve them, are examined further in section 5.

4.2. Gaps Caused by Cloud Coverage

[31] One important requirement for the direct sampling is for variables to be reasonably spatially stationary. In other words, the known locations/dates sampled for reconstruction should be statistically representative of the unknown areas. Spatial stationarity is reasonably honored in the case of gaps caused by satellite orbital characteristics because the locations and dates of the missing data are independent of the observed values. Gaps are equally likely to occur for high or low values of the variable being reconstructed, as the satellite does not selectively erase data. However, the situation is potentially different with gaps caused by cloud coverage, as a direct physical response is introduced into the already nonlinear system. In this case, gaps that are located under cloudy areas are likely to have lower latent heat fluxes and lower temperatures than those locations not obscured by clouds as a result of energy balance considerations. Likewise, while the soil moisture will not be as affected by clouds of short duration, if these are precipitating, there is the expectation that near surface moisture will increase relative to the noncloudy regions. As a result, if reconstruction of one variable is based on another that has undergone a physical response during the sampling procedure, it is likely

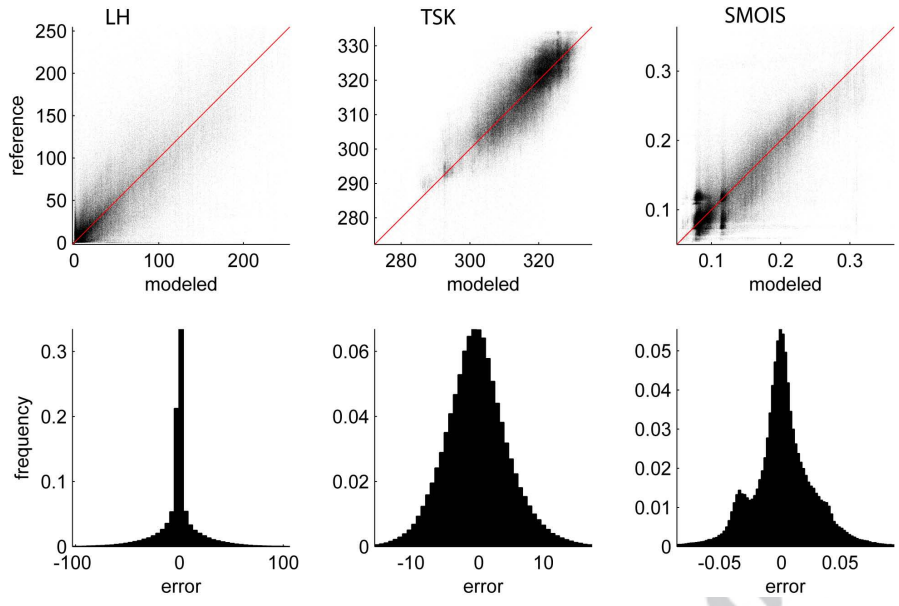

Figure 3. Gaps resulting from imposition of orbital characteristics. (top) Scatterplots and (bottom) errors in the three reconstructed variables of (from left to right) latent heat (LH), surface temperature (TSK), and soil moisture (SMOIS) for all reconstructed values in the month of January 2006.

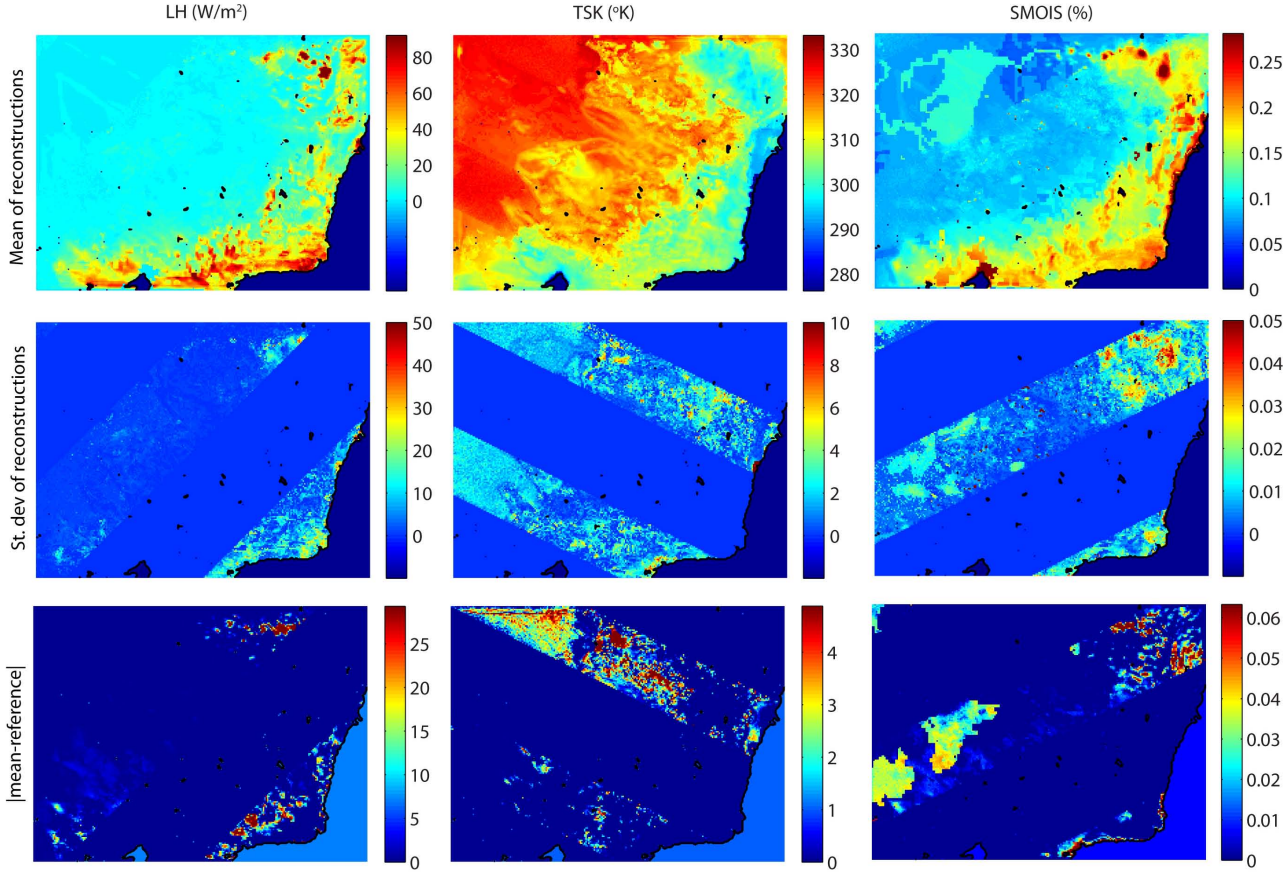

Figure 4. Statistics on the reconstruction ensembles for 15 January 2006. The three columns represent WRF simulated data fields of (from left to right) latent heat flux (LH), surface temperature (TSK), and soil moisture (SMOIS). The rows (from top to bottom) describe the ensemble mean of the reconstructed fields, the standard deviation of ensembles of the reconstructed images, and the ensemble mean minus the original WRF simulation.

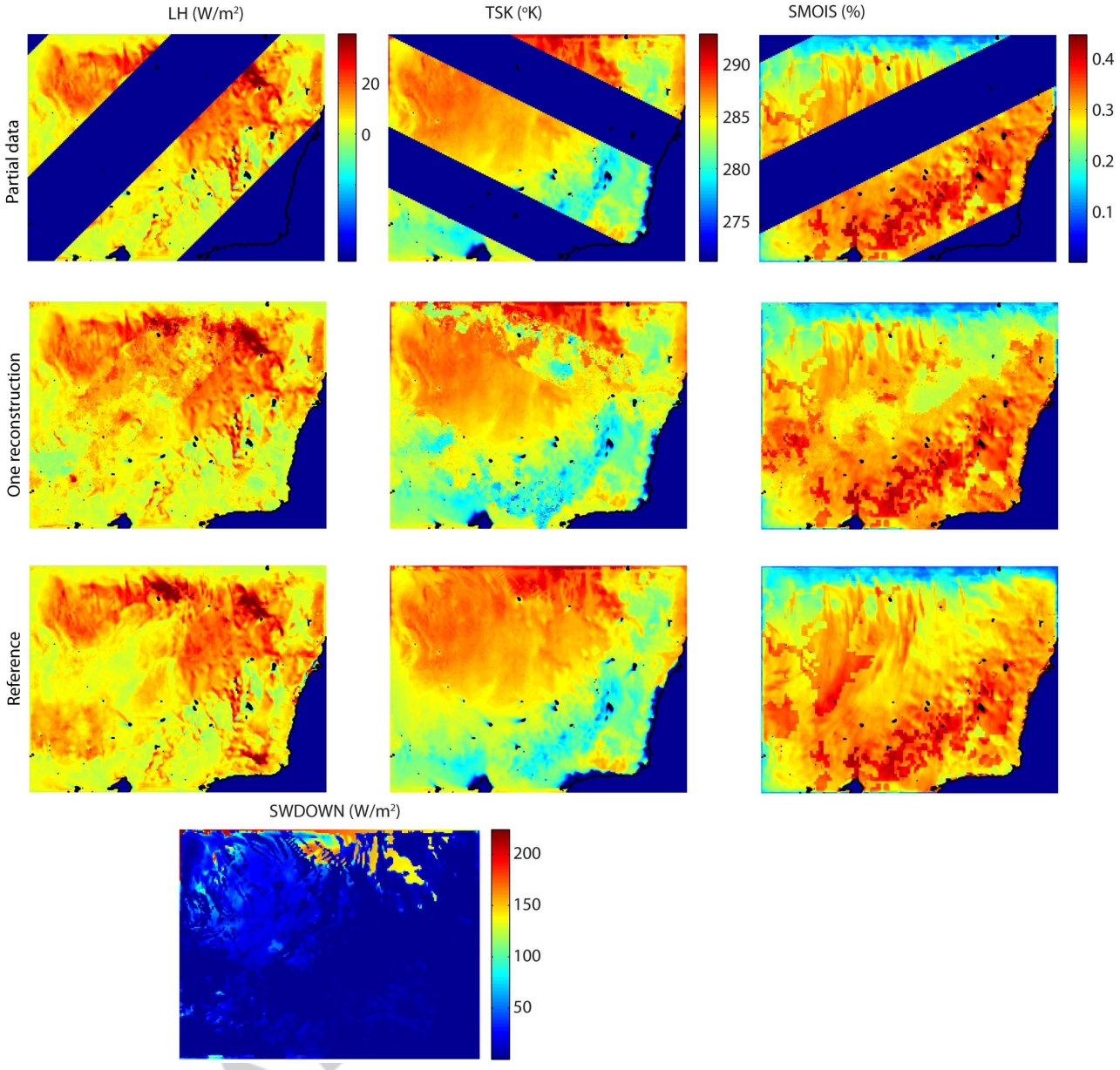

Figure 5. Gaps caused by orbital characteristics in a sample reconstruction for 15 July 2006. As in Figure 3, the three columns representing WRF simulated data fields are (from left to right) latent heat flux (LH), surface temperature (TSK), and soil moisture (SMOIS). The rows detail (from top to bottom) the artificially gap enforced simulation, the reconstructed image, and the original continuous WRF simulation. Downward shortwave radiation is included at the bottom of the panel.

that additional errors will be introduced into the reconstruction process.

[32] This bias has been observed in numerical experiments similar to those undertaken in section 4.1, but with artificial gaps made by removing all values where the shortwave downward radiation is less than 250 W m^{-2} (an arbitrary threshold representing likely cloud cover during summer). Figure 9 shows the reconstruction results with the cloud coverage gaps from 1 January 2006. It should be noted that there is a combination of gap sizes, with small gaps scattered throughout the domain and a large missing portion in the southwest of the domain. In the small gaps, the reconstruction accurately reproduces the reference patterns. This successful reconstruction is possible because nearby informed values constitute strong constraints on the type of patterns to use in the gaps. In the large missing area however, the reconstructed values are systematically biased, as expected, resulting in higher surface temperatures and lower soil moistures than the reference (which models the inherent land-atmosphere feedback relationship through WRF). This is also visible in the histograms of errors (Figure 10), where the errors are not centered on zero for temperature and soil moisture. At least in this example, it seems that our method is subject to the above described bias when filling large gaps caused by cloud cover (weakly constrained problem). When the gaps are small compared to the size of the spatial structures, it

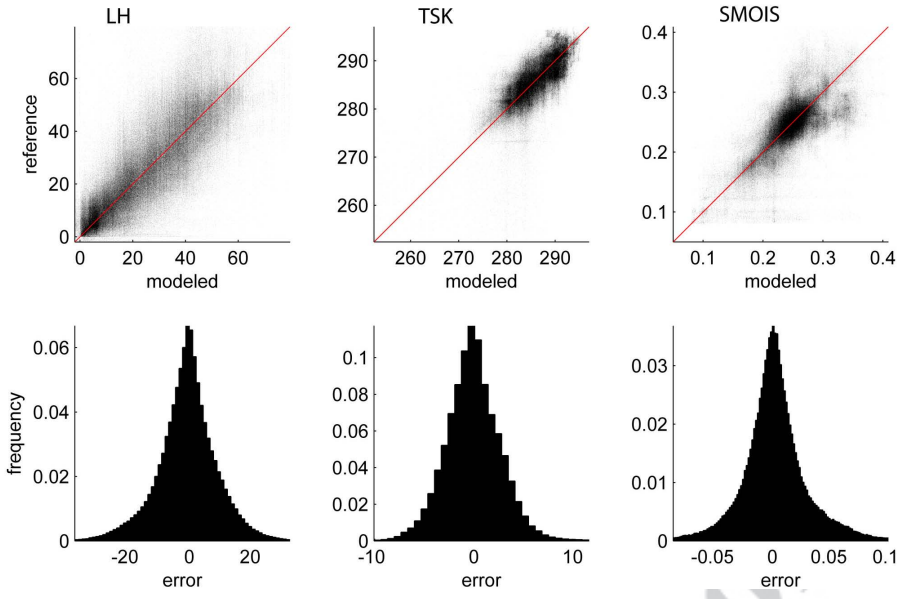

Figure 6. Gaps resulting from imposition of orbital characteristics. (top) Scatterplots and (bottom) errors in the three reconstructed variables of (from left to right) latent heat (LH), surface temperature (TSK), and soil moisture (SMOIS) for all reconstructed values in the month of July 2006.

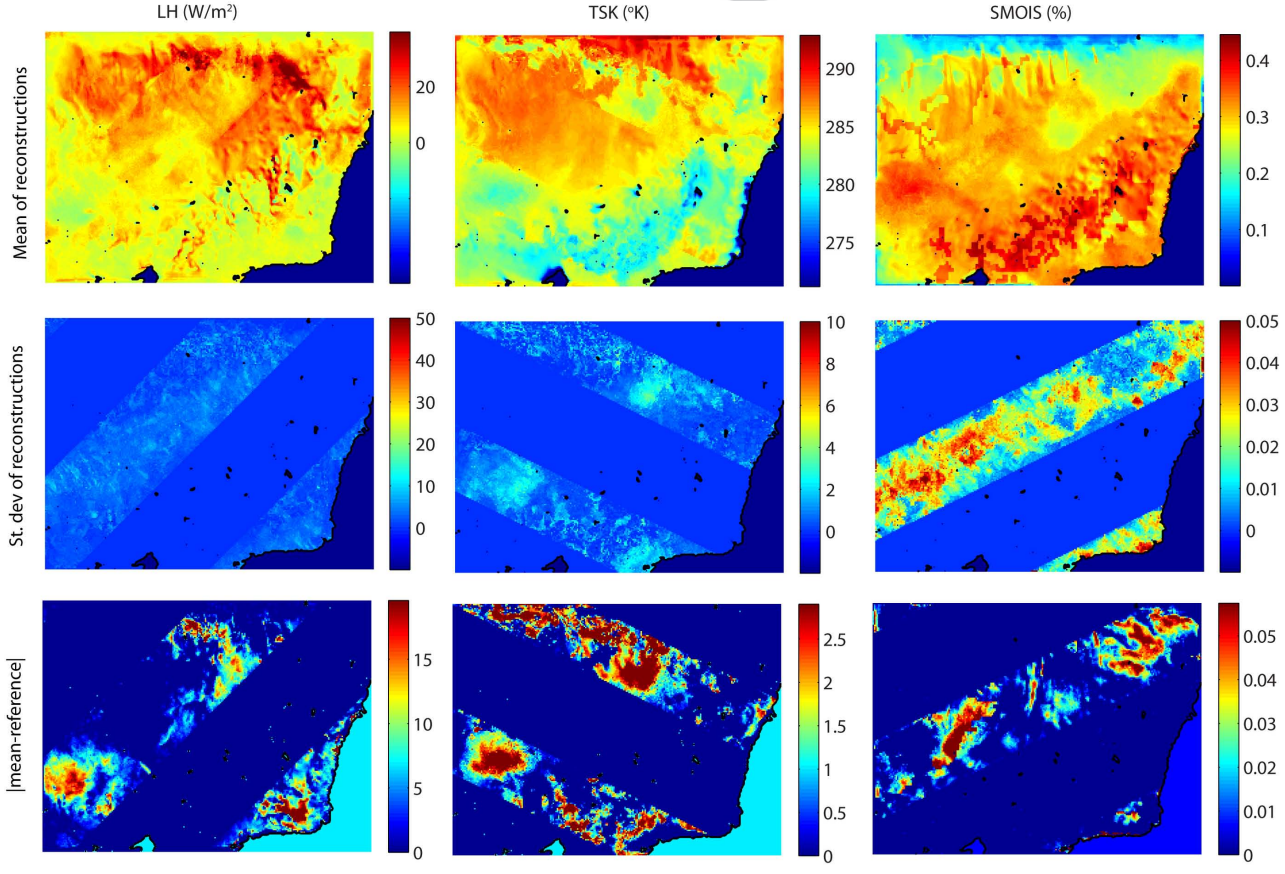

Figure 7. Statistics on reconstruction ensemble for 15 July 2006. As in Figure 5, the three columns represent WRF simulated data fields of (from left to right) latent heat flux (LH), surface temperature (TSK), and soil moisture (SMOIS). The rows (from top to bottom) describe the ensemble mean of the reconstructed fields, the standard deviation of ensembles of reconstructed image, and the ensemble mean minus the original WRF simulation.

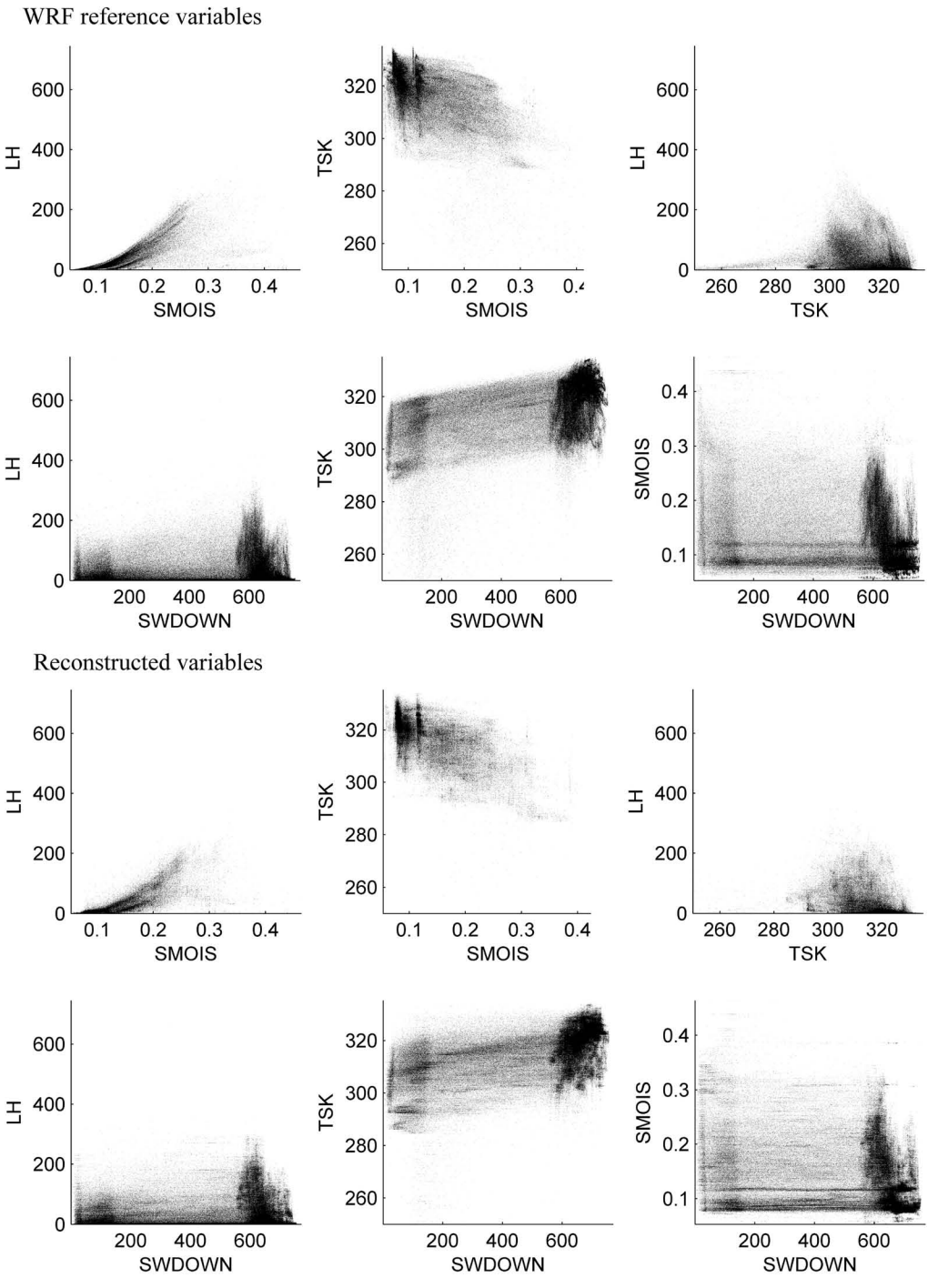

Figure 8. Validation of the multivariate joint distributions. The first and second rows are sample scatterplots of the WRF reference variables for all dates in January, illustrating varying degrees of nonlinear relationships between variables in the simulations. The third and fourth rows are scatterplots of the reconstructed variables for all dates in January, exhibiting good visual agreement of the nonlinear relationships exhibited in the WRF reference data.

seems that the gap-filling problem is constrained enough to overcome this bias, resulting in visually satisfying results.

5. Discussion and Conclusion

[33] Realistic reconstruction of missing data in remote sensing retrievals is a challenging problem that goes far beyond simple interpolation. Here we present a newly developed geostatistical approach that accounts for the complex spatial, structural and textural properties of the variables considered and the inherent nonlinear relationships in Earth surface variables. The method is based on a conditional stochastic resampling of the known areas of the domain, at different locations and dates. We assess the

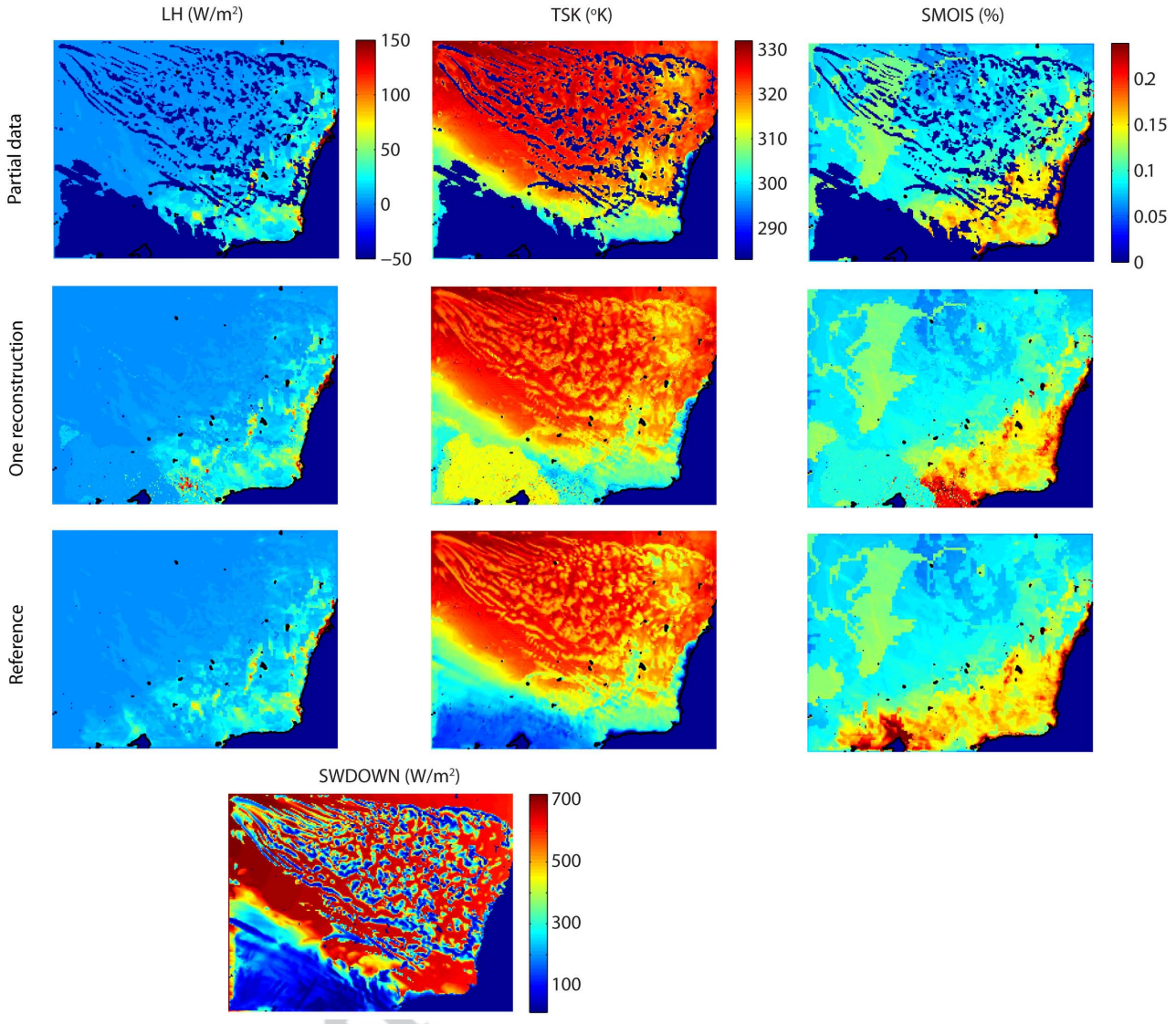

Figure 9. A sample reconstruction for 1 January 2006 using the SWDOWN image and a threshold of 250 W m^{-2} to identify cloud cover. Three columns representing WRF simulated data fields are (from left to right) latent heat flux (LH), surface temperature (TSK), and soil moisture (SMOIS). The rows detail (from top to bottom) the artificially gap enforced simulation, the reconstructed image, and the original continuous WRF simulation. Downward shortwave radiation is included at the bottom.

potential of the method in reconstructing gaps that would occur operationally in satellite observations by using synthetic data from a regional climate model of southeastern Australia, with both atmospheric and exaggerated scan gaps covering large parts of the domain.

[34] A question that is not addressed in most reconstruction studies is the uncertainty related to the interpolation results. It should be emphasized here that any interpolation entails uncertainty at the reconstructed locations, and this uncertainty is dependent of the level of spatiotemporal variability of the phenomenon considered. Variables that respond slowly in time, such as soil moisture or vegetation cover, can be very accurately recovered from measurements because the historic data contain high localized information content. From year to year, the same areas tend to have similar soil moisture or leaf area index [*Fang et al.*, 2008; *Yuan et al.*, 2011]. For example, *Wang et al.* [2012] accurately fill gaps in daily soil moisture remote sensing measurements based on known soil moisture taken one or a few days before. This is made possible because at the scale considered (0.5 degree resolution) daily variations of soil moisture are relatively small. Hence the data of the previous day strongly conditions the gap-filling problem and the uncertainty in the results is minimal. Similarly, *Boloorani et al.* [2008] fill gaps in multispectral satellite images given an image of the same location a year earlier. The process considered presents very small temporal variability and therefore the available data are strongly informative, resulting in low uncertainty.

[35] Just as the temporal stability (or not) of a reconstructed variable is important, so is the degree of spatial heterogeneity inherent in the retrieval. Small gaps in a smooth image can be filled with realistic values because some structures will be present on either side of the gap and relatively simple deterministic interpolation methods

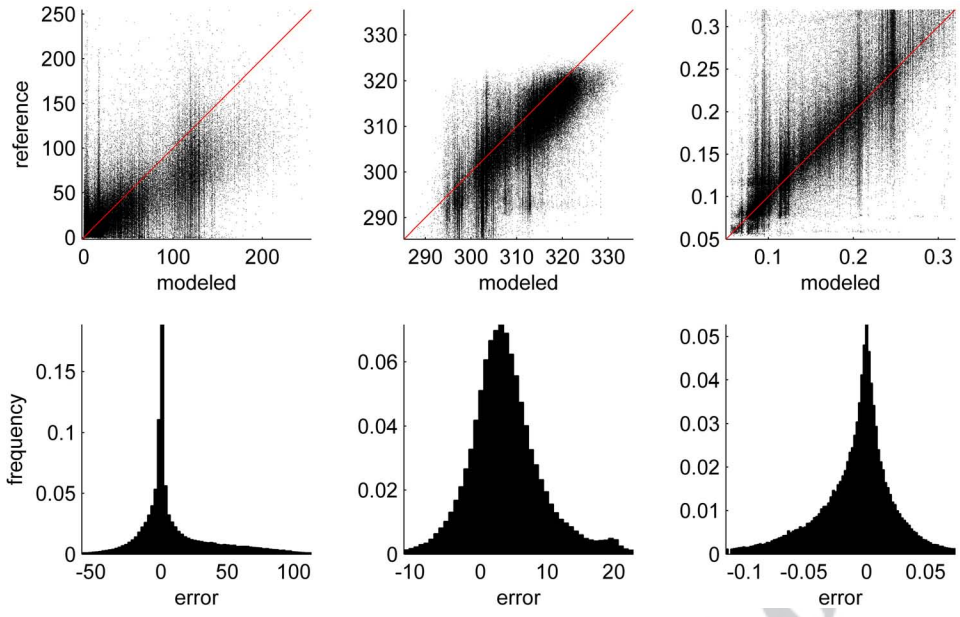

Figure 10. Gaps caused by cloud coverage with (top) scatterplots and (bottom) errors in the three reconstructed variables of (from left to right) latent heat (LH), surface temperature (TSK), and soil moisture (SMOIS) for all reconstructed values in the month of January 2006.

will give good results. For example, *Maxwell et al.* [2007] infills Landsat images where the gaps considered were small and the continuity of structures clearly visible between the gaps. Similarly, *Zhang et al.* [2007] use geostatistics for gap filling of remote sensing images with relatively small gaps. In that example, the kriging technique is used, which is known to be a smooth interpolator that cannot reproduce specific complex spatial features [*Journel and Zhang*, 2006; *Olea*, 1996]. An additional issue that is generally not addressed is the filling of gaps that simultaneously occult several interrelated variables. The additional constraint in this case is that the different synthetically generated values must present the correct (nonlinear) dependencies with each other.

[36] Results from this analysis show that our direct sampling based gap-filling method is able to realistically reconstruct the missing elements of the synthetic satellite images. Complex spatial patterns can be resolved that also reproduce the often nonlinear dependencies observed between the variables considered. Furthermore, the stochastic nature of the methodology makes it possible to ascertain the uncertainty related to the reconstruction. The governing principle is to use past occurrences of observed multivariate multiple-point relationships and then apply these to the present. This approach however assumes that the realm of possible outcomes is contained in the past observations, an assumption that may not always hold, explaining artifacts such as discontinuities or departure from the reference values (e.g., soil moisture for the January example in Figure 2 or temperature for the July example in Figure 5).

[37] The definition of a distance measure between patterns is necessary in the resampling procedure. Because multivariate patterns are considered, weights associated with the different variables need to be defined. Although in this study we assigned identical values to all weights (i.e., all weights $= 1/m$), one could consider modifying these to give more importance to the variable being simulated, relative to the other variables. This could potentially be used in contexts where certain variables are acquired with higher confidence than others. One aspect is that reducing the relative weight of the other variables could potentially result in a decreased reproduction of the multivariate relationships. When the information added by ancillary variables is not in perfect agreement with the observed spatial continuity, choices have to be made to honor either one or the other of the variables being used. The assigned weights would then express a necessary trade-off between these constraints. Weights might also be varied locally, or their computation could be based on the relative information content of each variable. Investigating the use of importance weights to individual variables is an area that requires further investigation and refinement for different applications.

[38] The best results are obtained with gaps caused by satellite orbital characteristics, because the locations of the missing data are independent of the observed values. In the case of gaps caused by cloud coverage, the gaps preferentially occur at locations of higher than average soil moistures and lower than average temperatures. The consequence is that the statistics of the known portions of the domain do not correspond to the statistics of the target locations to reconstruct, resulting in a statistical bias that is especially pronounced for large gaps. A way of overcoming this bias would be to infer the multivariate high-order statistics not from the known portions of the domain, which are preferentially sampled, but from regional climatic models (RCMs) that provide exhaustive coverage, including the areas covered by clouds. The methodology would not change, but the fundamental challenge would then be to obtain a RCM that reproduces the spatial patterns observed in real remote sensing images, including structural and measurement noise. If fully informed training images were available, another application could also be the stochastic generation of spatially distributed weather variables.

[39] **Acknowledgments.** This work was supported by research projects undertaken as part of the Australian Research Council and National Water Commission funding for the National Centre for Groundwater Research and Training (NCGRT). We thank Jason Evans from the UNSW Climate Change Research Centre for access to the WRF simulations. We would also like to thank the three anonymous reviewers, whose comments helped to improve the final paper.

References

Boloorani, A. D., S. Erasmi, and M. Kappas (2008), Multi-source remotely sensed data combination: Projection transformation gap-fill procedure, *Sensors*, *8*(7), 4429–4440.

Brunner, P., P. Bauer, M. Eugster, and W. Kinzelbach (2004), Using remote sensing to regionalize local precipitation recharge rates obtained from the chloride method, *J. Hydrol.*, *294*(4), 241–250, doi:10.1016/j.jhydrol.2004.02.023.

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

Chilès, J.-P., and P. Delfiner (1999), *Geostatistics: Modeling Spatial Uncertainty*, John Wiley, New York.

Cihlar, J. (2000), Land cover mapping of large areas from satellites: Status and research priorities, *Int. J. Remote Sens.*, *21*(6–7), 1093–1114.

Diak, G. R., W. L. Bland, and J. Mecikalski (1996), A note on first estimates of surface insolation from GOES-8 visible satellite data, *Agric. For. Meteorol.*, *82*, 219–226.

Drusch, M., E. F. Wood, H. Gao, and A. Thiele (2004), Soil moisture retrieval during the Southern Great Plains Hydrology Experiment 1999: A comparison between experimental remote sensing data and operational products, *Water Resour. Res.*, *40*, W02504, doi:10.1029/2003WR002441.

Du, Y., J. Cihlar, J. Beaubien, and R. Latifovic (2001), Radiometric normalization, compositing, and quality control for satellite high resolution image mosaics over large areas, *IEEE Trans. Geosci. Remote Sens.*, *39*(3), 623–634.

Evans, J. P., and M. F. McCabe (2010), Regional climate simulation over Australia's Murray-Darling basin: A multitemporal assessment, *J. Geophys. Res.*, *115*, D14114, doi:10.1029/2010JD013816.

Fang, H., S. Liang, J. R. Townshend, and R. E. Dickinson (2008), Spatially and temporally continuous LAI data sets based on an integrated filtering method: Examples from North America, *Remote Sens. Environ.*, *112*(1), 75–93.

Ferguson, C. R., and E. F. Wood (2010), An evaluation of satellite remote sensing data products for land surface hydrology: Atmospheric infrared sounder, *J. Hydrometeorol.*, *11*(6), 1234–1262.

Goovaerts, P. (1997), *Geostatistics for Natural Resources Evaluation*, Oxford Univ. Press, Oxford, U. K.

Goovaerts, P. (2000), Geostatistical approaches for incorporating elevation into the spatial interpolation of rainfall, *J. Hydrol.*, *228*(1–2), 113–129.

Huffman, G. J., R. F. Adler, B. Rudolph, U. Schneider, and P. Keehn (1995), Global precipitation estimates based on a technique for combining satellite-based estimates, rain gauge analysis, and NWP model precipitation information, *J. Clim.*, *8*, 1284–1295.

Jeu, R. A. M., W. Wagner, T. R. H. Holmes, A. J. Dolman, N. C. Giesen, and J. Friesen (2008), Global soil moisture patterns observed by space borne microwave radiometers and scatterometers, *Surv. Geophys.*, *29*(4–5), 399–420.

Journel, A., and T. Zhang (2006), The necessity of a multiple-point prior model, *Math. Geol.*, *38*(5), 591–610.

Kalma, J. D., T. R. McVicar, and M. F. McCabe (2008), Estimating land surface evaporation: A review of methods using remotely sensed surface temperature data, *Surv. Geophys*, *29*(4–5), 421–469.

Kummerow, C., J. Simpson, O. Thiele, J. Barnes, A. T. C. Chang, E. Stocker, R. F. Adler, and A. Hou (2000), The status of the Tropical Rainfall Measuring Mission (TRMM) after two years in orbit, *J. Appl. Meteorol.*, *39*(12), 1965–1982.

Kustas, W. P., and J. M. Norman (2000), Evaluating the effects of subpixel heterogeneity on pixel average fluxes, *Remote Sens. Environ.*, *74*, 327–342.

Leuangthong, O., and C. Deutsch (2003), Stepwise conditional transformation for simulation of multiple variables, *Math. Geol.*, *35*(2), 155–173.

Li, H. T., P. Brunner, W. Kinzelbach, W. P. Li, and X. G. Dong (2009), Calibration of a groundwater model using pattern information from remote sensing data, *J. Hydrol.*, *377*(1–2), 120–130.

Liu, Y. Y., R. M. Parinussa, W. A. Dorigo, R. A. M. De Jeu, W. Wagner, A. I. J. M. Van Dijk, M. F. McCabe, and J. P. Evans (2011), Developing an improved soil moisture dataset by blending passive and active microwave satellite-based retrievals, *Hydrol. Earth Syst. Sci.*, *15*(2), 425–436.

Mariethoz, G., and B. F. J. Kelly (2011), Modeling complex geological structures with elementary training images and transform-invariant distances, *Water Resour. Res.*, *47*, W07527, doi:10.1029/2011WR010412.

Mariethoz, G., and P. Renard (2010), Reconstruction of incomplete data sets or images using direct sampling, *Math. Geosci.*, *42*(3), 245–268, doi:10.1007/s11004-010-9270-0.

Mariethoz, G., P. Renard, and R. Froidevaux (2009), Integrating collocated auxiliary parameters in geostatistical simulations using joint probability distributions and probability aggregation, *Water Resour. Res.*, *45*, W08421, doi:10.1029/2008WR007408.

Mariethoz, G., P. Renard, and J. Straubhaar (2010), The direct sampling method to perform multiple-point geostatistical simulations, *Water Resour. Res.*, *46*, W11536, doi:10.1029/2008WR007621.

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

McCabe, M. F., H. Gao, and E. F. Wood (2005), Evaluation of AMSR-E-derived soil moisture retrievals using ground-based and PSR airborne data during SMEX02, *J. Hydrometeorol.*, *6*(6), 864–877.

McCabe, M., E. Wood, R. Wojcik, M. Pan, J. Sheffield, H. Gao, and H. Su (2008a), Hydrological consistency using multi-sensor remote sensing data for water and energy cycle studies, *Remote Sens. Environ.*, *112*(2), 430–444.

McCabe, M. F., L. K. Balick, J. Theiler, A. R. Gillespie, and A. Mushkin (2008b), Linear mixing in thermal infrared temperature retrieval, *Int. J. Remote Sens.*, *29*(17–18), 5047–5061.

Milzow, C., L. Kgotlhang, W. Kinzelbach, P. Meier, and P. Bauer-Gottwein (2009), The role of remote sensing in hydrological modelling of the Okavango Delta, Botswana, *J. Environ. Manage.*, *90*(7), 2252–2260.

Olea, R. A. (1996), Compensating for estimation smoothing in kriging, *Math. Geol.*, *28*(4), 407–417.

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

Rivoirard, J. (2001), Which models for collocated cokriging, *Math. Geol.*, *33*(2), 117–131.

Sahoo, A. K., M. Pan, T. J. Troy, R. K. Vinukollu, J. Sheffield, and E. F. Wood (2011), Reconciling the global terrestrial water budget using satellite remote sensing, *Remote Sens. Environ.*, *115*(8), 1850–1865.

Shannon, C. (1948), A mathematical theory of communication, *Bell Syst. Tech. J.*, *27*, 379–423.

Sheffield, J., C. R. Ferguson, T. J. Troy, E. F. Wood, and M. F. McCabe (2009), Closing the terrestrial water budget from satellite remote sensing, *Geophys. Res. Lett.*, *36*, L07403, doi:10.1029/2009GL037338.

Su, H., E. F. Wood, M. F. McCabe, and Z. Su (2007), Evaluation of remotely sensed evapotranspiration over the CEOP EOP-1 reference sites, *J. Meteorol. Soc. Jpn.*, *85A*, 439–459.

Wan, Z., Y. Zhang, Q. Zhang, and Z.-L. Li (2002), Validation of the land-surface temperature products retrieved from Terra Moderate Resolution Imaging Spectroradiometer data, *Remote Sens. Environ.*, *83*, 163–180.

Wang, G., D. Garcia, Y. Lui, R. de Jeu, and A. Dolman (2012), A three-dimensional gap filling method for large geophysical datasets: Application to global satellite soil moisture observations, *Environ. Modell. Software*, *30*, 139–142.

Weymouth, G. T., and J. F. Le Marshall (2001), Estimation of daily surface solar exposure using GMS-5 stretched-VISSR observations: The system and basic results, *Aust. Meteorol. Mag.*, *50*(4), 263–278.

Yeh, J., M. Jin, and S. Hanna (1996), An iterative stochastic inverse method: Conditional effective transmissivity and hydraulic head fields, *Water Resour. Res.*, *32*(1), 85–92.

Yuan, H., Y. Dai, Z. Xiao, D. Ji, and W. Shangguan (2011), Reprocessing the MODIS leaf area index products for land surface and climate modelling, *Remote Sens. Environ.*, *115*(5), 1171–1187.

Zhang, C., W. Li, and D. Travis (2007), Gaps-fill of SLC-off Landsat ETM+ satellite image using a geostatistical approach, *Int. J. Remote Sens.*, *28*(22), 5103–5122.

Zhang, C., W. Li, and D. J. Travis (2009), Restoration of clouded pixels in multispectral remotely sensed imagery with cokriging, *Int. J. Remote Sens.*, *30*(9), 2173–2195.

Zhang, X., H. Jiang, G. Zhou, Z. Xiao, and Z. Zhang (2012), Geostatistical interpolation of missing data and downscaling of spatial resolution for remotely sensed atmospheric methane column concentrations, *Int. J. Remote Sens.*, *33*(1), 120–134.

Author Queries

AQ1: Please verify that "sections 2 and 3" was meant here, or edit further if necessary.