Mixtures of Gaussians for Uncertainty Description in Bivariate Latent Heat Flux Proxies

R. WÓJCIK,* PETER A. TROCH, H. STRICKER, AND P. TORFS

Environmental Sciences Group, Hydrology and Quantitative Water Management, Wageningen University, Wageningen, Netherlands

E. WOOD AND H. SU

Department of Civil and Environmental Engineering, Princeton University, Princeton, New Jersey

Z. SU

International Institute for Geo-Information Science and Earth Observation (ITC), Enschede, Netherlands

(Manuscript received 23 February 2005, in final form 9 September 2005)

ABSTRACT

This paper proposes a new probabilistic approach for describing uncertainty in the ensembles of latent heat flux proxies. The proxies are obtained from hourly Bowen ratio and satellite-derived measurements, respectively, at several locations in the southern Great Plains region in the United States. The novelty of the presented approach is that the proxies are not considered separately, but as bivariate samples from an underlying probability density function. To describe the latter, the use of Gaussian mixture density models—a class of nonparametric, data-adaptive probability density functions—is proposed. In this way any subjective assumptions (e.g., Gaussianity) on the form of bivariate latent heat flux ensembles are avoided. This makes the estimated mixtures potentially useful in nonlinear interpolation and nonlinear probabilistic data assimilation of noisy latent heat flux measurements. The results in this study show that both of these applications are feasible through regionalization of estimated mixture densities. The regionalization scheme investigated here utilizes land cover and vegetation fraction as discriminatory variables.

1. Introduction

Latent heat flux (LE) is the key variable that provides a link between energy and water budgets at the land surface. Since much of our understanding of the complex feedback mechanisms between the earth surface and the atmosphere is focused on quantifying these budgets, there is considerable interest in developing methods that routinely predict this variable. Local- and regional-scale estimates of LE would offer insight into hydroecological processes, aid in improving irrigation efficiency, and would provide a valuable tool for water resource management. Accurate estimation at large scales is required to improve our understanding of the global climate and its spatial and temporal variability (Miller et al. 1995). However, the prediction and validation of LE across all scales remains problematic.

The conventional methods to estimate LE are based on point measurements of energy balance components or turbulent surface fluxes and are representative only for very local scales. Recently, a new class of techniques based on satellite remotely sensed (RS) information has been developed to compute LE at scales from 1 km to a continent. Despite their theoretical attractiveness, especially for regional and global hydrological applications, "satellite derived" LE_{sat} usually does not compare well with "in situ measured" LE_{is}. Both proxies of LE, however, contain information about the true variability of this quantity. The difficulty in inferring this information from data is due to different sources of uncertainty involved (e.g., measurement errors, support scale, heterogeneity of land surface). In this context it is

* Current affiliation: Department of Civil and Environmental Engineering, Princeton University, Princeton, New Jersey.

Corresponding author address: R. Wójcik, Dept. of Civil and Environmental Engineering, Princeton University, Princeton, NJ 08544.
E-mail: rwojcik@princeton.edu

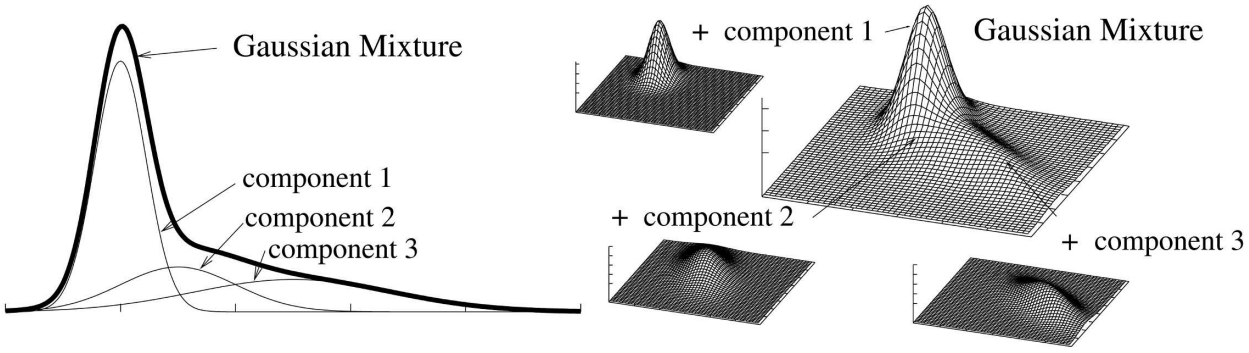

FIG. 1. 1D and 2D example of MG (in both cases as a linear combination of three components).

therefore natural to treat LE as a joint probability density function (pdf) over a spatially distributed set of bivariate random variables comprised of both proxies. Although this "true" high-dimensional joint pdf is a purely theoretical object, we are able to observe the bivariate samples from its marginal pdfs at particular locations in space. The purpose of this paper is threefold: to investigate a new nonparametric methodology for fitting these marginal LE pdfs to the experimental data from six sites, to pose the (preliminary) hypothesis of regionalization of the estimated pdfs through land use alone and additional parameters like degree of vegetation cover using again the six datasets, and to show the theoretical option of using these pdfs in spatiotemporal interpolation and data assimilation. Technically, the last objective is accomplished by first estimating the marginal pdf of bivariate LE and then using it to derive a conditional pdf of LE_{is} given LE_{sat}. The motivation for the latter asymmetry is that since in situ measurements of LE are derived from observations of physical processes at the land surface that determine natural variability of LE, they provide best estimates of LE at local scale. On the other hand, satellites observe states that are affecting this flux (as, e.g., surface temperature) at coarser pixel scales. Moreover, when RS information is used to derive LE there are extra sources of uncertainty as compared to in situ measurements due to inadequacies in retrieval algorithms, nonlinear measurement error propagation through RS models for LE, influence of cloud cover, and errors in land-use classification (Hipps and Kustas 2000). In this paper we therefore choose to condition local LE_{is} on LE_{sat}. The conditionals are modeled by a *nonparametric* class of continuous pdfs referred to as mixtures of Gaussians (MGs; see McLachlan and Peel 2000). The attractive property of MGs is that they do not require any arbitrary assumptions on the form of an underlying pdf (like, e.g., Gaussian assumption). This implies that as compared to the classical parametric approaches MGs can adapt to the local geometry of data ensembles (e.g., points distributed in multiple modes or points distributed on a low-dimensional surface in a high-dimensional space) and are able to approximate any continuous density to an arbitrary precision. Moreover, it is easy to simulate ensembles of points from parameterized MGs, which makes them useful in ensemble Kalman filtering.

The marginal MGs can further be regionalized and used for spatiotemporal interpolation of the LE proxies. In this article we investigate a practical method for regionalization of MGs that utilizes land use and vegetation cover as discriminatory variables. The interpolation is performed by deriving the conditional pdfs from a particular regionalized marginal pdf and then calculating their (conditional) expectation. Such nonlinear interpolation is particularly suitable for the LE flux, which does not aggregate/disaggregate linearly (Braud 1998). Another benefit from having the regionalized conditionals parameterized by MGs is that they can be assimilated into land surface models by the recently developed nonlinear ensemble Kalman filter (see Anderson and Anderson 1999; Torfs et al. 2002).

The LE_{is} data in this paper are obtained from energy balance Bowen ratio (EBBR) systems from southern Great Plains (SGP) region in the United States. The LE_{sat} proxies are estimated with Surface Energy Balance System (SEBS) developed by Su (2002).

2. Mixtures of Gaussians

a. Definition

To estimate the conditional uncertainty of LE_{is} given LE_{sat} a pdf needs first to be fitted to a bivariate sample $\{LE_{sat,k}; LE_{is,k}\}_{k=1}^{K}$. In this work the focus is on the use of MGs, which are defined as a linear combination of Gaussian densities (see Fig. 1), called components:

$$p(x_1, \ldots, x_D) = p(\mathbf{x}) = \sum_{n=1}^{N_c} w_n g_{(\mathbf{m}_n, \mathbf{C}_n)}(\mathbf{x}), \quad (1)$$

where $\mathbf{x}$ is a D-dimensional vector of variables, N_c is the number of components, and $g_{(\mathbf{m}_n, \mathbf{C}_n)}$ stands for the Gaussian density with mean $\mathbf{m}_n$ and covariance $\mathbf{C}_n$.

Here $D = 2$, $\mathbf{x} = [\mathrm{LE_{is}}, \mathrm{LE_{sat}}]^{\mathrm{T}}$ and the w_n's are the component weights that "$\forall_n\ w_n \geq 0$" and $\Sigma w_n = 1$. Densities of the MG type inherit a lot of interesting properties from their Gaussian components: for example, the conditional densities $p(x_1|x_2)$,

$$p(x_1|x_2) = \frac{p(x_1,x_2)}{\int d\eta p(x_2,\eta)}, \quad (2)$$

are again MGs and can be calculated analytically (for technical details see Sharma 2000; Torfs and Wójcik 2001), Monte Carlo sampling is easy and fast, and, if needed, the conditional expectation (regression curve),

$$E[x_1|x_2] = \int dx_1 x_1 p(x_1|x_2), \quad (3)$$

can be derived from (1) and again is given by an analytic expression (Torfs and Wójcik 2001).

b. Fitting procedure

Fitting (1) to data requires optimizing weights w_n and the parameters of the components $\theta_{\mathbf{n}} = \{\mathbf{m}_n, \mathbf{C}_n\}$. The most commonly used optimization criterion is the maximum likelihood (ML) criterion implemented as expectation maximization (EM) algorithm (McLachlan and Krishnan 1997). Standard EM for mixtures, however, exhibits some weaknesses. It requires knowledge of N_c and good initialization is essential for reaching a good local optimum. To overcome these difficulties we use the approach of Figueiredo and Jain (2002) based on the minimum message length (MML) criterion. The rationale behind MML is that if one can built a short code describing one's data that means that one has a good data generation model (Bishop 1995). Mathematically, the MML criterion for MG pdfs consists of minimizing with respect to θ, where $\theta \equiv \{\theta_1 \ldots \theta_{N_c}, w_1 \ldots w_{N_c}\}$, the following cost function:

$$\mathcal{L}(\theta,\mathbf{x}) = \frac{N}{2} \sum_{n=1}^{N_c} \log\left(\frac{K w_n}{12}\right) + \frac{N_c}{2} \log\left(\frac{K}{12}\right) + \frac{N N_c}{2} - \log[p(\mathbf{x}|\theta)], \quad (4)$$

where $N = \dim(\theta_n)$, and K is the number of sample points. An attractive property of the algorithm of Figueiredo and Jain (2002) is that it is coupled with the model selection procedure that automatically determines the number of components N_c. Thus, MG can be initialized with a large value of N_c, alleviating the need for careful initialization. Because of this, a component-wise version of EM (Celeux et al. 2001) is adopted in Figueiredo and Jain (2002) to minimize (4).

3. Model structure of SEBS

The estimates of $\mathrm{LE_{sat}}$ were computed with SEBS (Su 2002). This model calculates atmospheric turbulent heat fluxes using satellite earth observation data. The SEBS consists of three components—land surface parameters, sensible heat flux estimation, and the energy balance—that are described briefly below [see Su (2002) for additional details].

Required land surface parameters include albedo, emissivity, surface temperature, fractional vegetation coverage, leaf area index, and the height of the vegetation from which displacement height and roughness height are derived. All this information is usually derived from remote sensing radiance data [e.g., the Moderate Resolution Imaging Spectroradiometer (MODIS)] in conjunction with other surface-related data [as, e.g., those from the Land Data Assimilation System (LDAS) database; http://ldas.gsfc.nasa.gov/LDAS8th/MAPPED.VEG/LDASmapveg.shtml]. The sensible heat flux H is based on the aerodynamic profile method and, because of the use of surface temperature, the determination of the two roughness lengths for heat and momentum transfer, as described in (Su et al. 2001). Required observations [or from four-dimensional data assimilation (4DDA) analysis fields] include air pressure, air temperature, humidity, and wind speed at a reference height (the measurement height for local-scale applications). For the local-scale SEBS obtains the friction velocity, the sensible heat flux, and the Monin–Obukhov stability length by solving iteratively a system of nonlinear equations. For field measurements performed at a height of a few meters above ground, where the surface fluxes are related to surface variables and variables in the atmospheric surface layer, all calculations involve the Monin–Obukhov similarity (MOS) functions given by Brutsaert (1999). The fluxes are based on the energy balance relations and utilize measured net radiation (or its equivalent derived through satellite-based observations of incoming radiation and surface temperature) and an estimate of the ground heat flux (see Su 2002). Latent heat flux can be estimated now from the energy balance equation:

$$\mathrm{LE_{sat}} = (R_n - G) - H_{\mathrm{sat}}, \quad (5)$$

where R_n stands for net radiation and G for soil heat flux, respectively.

4. MGs in the generalized ensemble Kalman filter

Apart from considering the conditionals in (2) as pure uncertainty descriptors, they can be assimilated into land surface models [as e.g., variable infiltration capacity (VIC) model in Liang et al. 1996] by the generalized ensemble Kalman filter (GEnKF). This algorithm [originally inspired by Anderson and Anderson (1999) and cast by Torfs et al. (2002) into a broader probability theoretical framework] goes beyond the classical linear ensemble Kalman filter (EnKF; Evensen 1994) in the sense that it does not require any a priori assumptions on the form of pdfs for state and observational noise, nor does it presumes linearity of state and/or output equations to get optimal state estimates. Accordingly, when new observations become available the state updates are not restricted to assimilating Gaussian pdfs of these observations as in the EnKF, but allow any nonparametric pdfs (e.g., MGs) to be incorporated. This new idea extends the work of Torfs et al. (2002) and makes GEnKF competitive with more traditional data assimilation schemes. Since the overall objective of this research direction is the implementation of GEnKF we find it useful here to give a brief mathematical description of the algorithm with emphasis on state updates given new observations.

Let us denote the state of the system at time n as s_n, the state at time $n + 1$ as s_{n+1}, and the observation at time $n + 1$ as o_{n+1}. These three variables can be scalars or vectors. We use ϕ_n to denote the joint pdf and f_n, f_{n+1} and h_{n+1} their respective marginals. Given a new observation o_{n+1} the solution to GEnKF problem is given by

$$f^g_{n+1}(s_{n+1}|o_{n+1}) = \frac{\int d\sigma_n \phi_n(\sigma_n, s_{n+1}, o_{n+1})}{\int d\sigma_n \int d\sigma_{n+1} \phi_n(\sigma_n, \sigma_{n+1}, o_{n+1})}. \tag{6}$$

Assuming conditional independence of s_n and o_{n+1} given s_{n+1} and making use of the Bayesian approach the following relations hold (Torfs et al. 2002):

$$f^p_{n+1}(s_{n+1}) = \int d\sigma_n f_n(\sigma_n) f_{n+1}(s_{n+1}|\sigma_n), \tag{7}$$

$$f^g_{n+1}(s_{n+1}|o_{n+1}) = \frac{h_{n+1}(o_{n+1}|s_{n+1}) f_{n+1}(s_{n+1})}{\int d\sigma_{n+1} h_{n+1}(o_{n+1}|\sigma_{n+1}) f_{n+1}(\sigma_{n+1})}. \tag{8}$$

Equation (7) is referred to as the *prediction* step and (8) as the *Kalman gain* or *analysis* step, respectively. When the observation is not known deterministically, but only its probability density φ is given,[1] this last formula is to be replaced by

$$F^g_{n+1}(s_{n+1}|\varphi) = \int d\omega_{n+1} \varphi(\omega_{n+1}) f^g_{n+1}(s_{n+1}|\omega_{n+1}). \tag{9}$$

With preliminary knowledge of $f_{n,n+1}(s_n, s_{n+1})$ and $\phi_{n+1}(s_{n+1}, o_{n+1})$, the integrals above are calculated recursively until the time of the latest observation.

Equations (6)–(9) describe an abstract setting for Kalman filtering, regardless of how the densities involved are given. In this paper we propose to approximate them by MGs. MGs are particularly well suited for this: because simulating from them is extremely fast, all integrals above can be evaluated by Monte Carlo sampling. Figure 2 illustrates the steps involved in computing the product in (9). First, a number of starting points is simulated from known observational φ_{n+1} pdf (Fig. 2a; the marks stand for the simulated points). From this ensemble, we calculate a statistically relevant set of posterior state pdfs f^g_{n+1} (on the lines of Fig. 2b). For this we use our knowledge of ϕ_{n+1} (Fig. 2c) at the simulated starting points. Then, the posteriors are again sampled, which is visualized in Fig. 2d. The joint pdf ϕ^*_{n+1} is then fitted in Fig. 2f to the sample in Fig. 2e. Finally, ϕ^*_{n+1} is marginalized (Fig. 2g), resulting in the analysis pdf F^g_{n+1}.

When dimensionality of state/observation space is high, as is the case of spatially distributed hydrologic models, the performance of the algorithm above would be suffering *the curse of dimensionality* (Gershenfeld 1992). That term was coined by Bellman (1961) to describe the problem that occurs when searching in or estimating pdfs on high-dimensional spaces. This problem may become intuitively clearer by looking at the example of fitting multidimensional histograms. Given a fixed number of M grid lines per dimension D, the number of independent cells grows as M^D. Furthermore, if the density function is to be estimated based on a set of high-dimensional samples, the number of samples required for accurate histogram estimation also grows as M^D. The same is true for MG models—here the components are continuous equivalents of cells used in histograming, and their weights can be viewed as histogram values at those cells. A pragmatic way to tackle this problem is to regionalize the pdfs. In

[1] For assimilation of LE into land surface models we propose $\varphi = p(\mathrm{LE_{is}}|\mathrm{LE_{sat}})$.

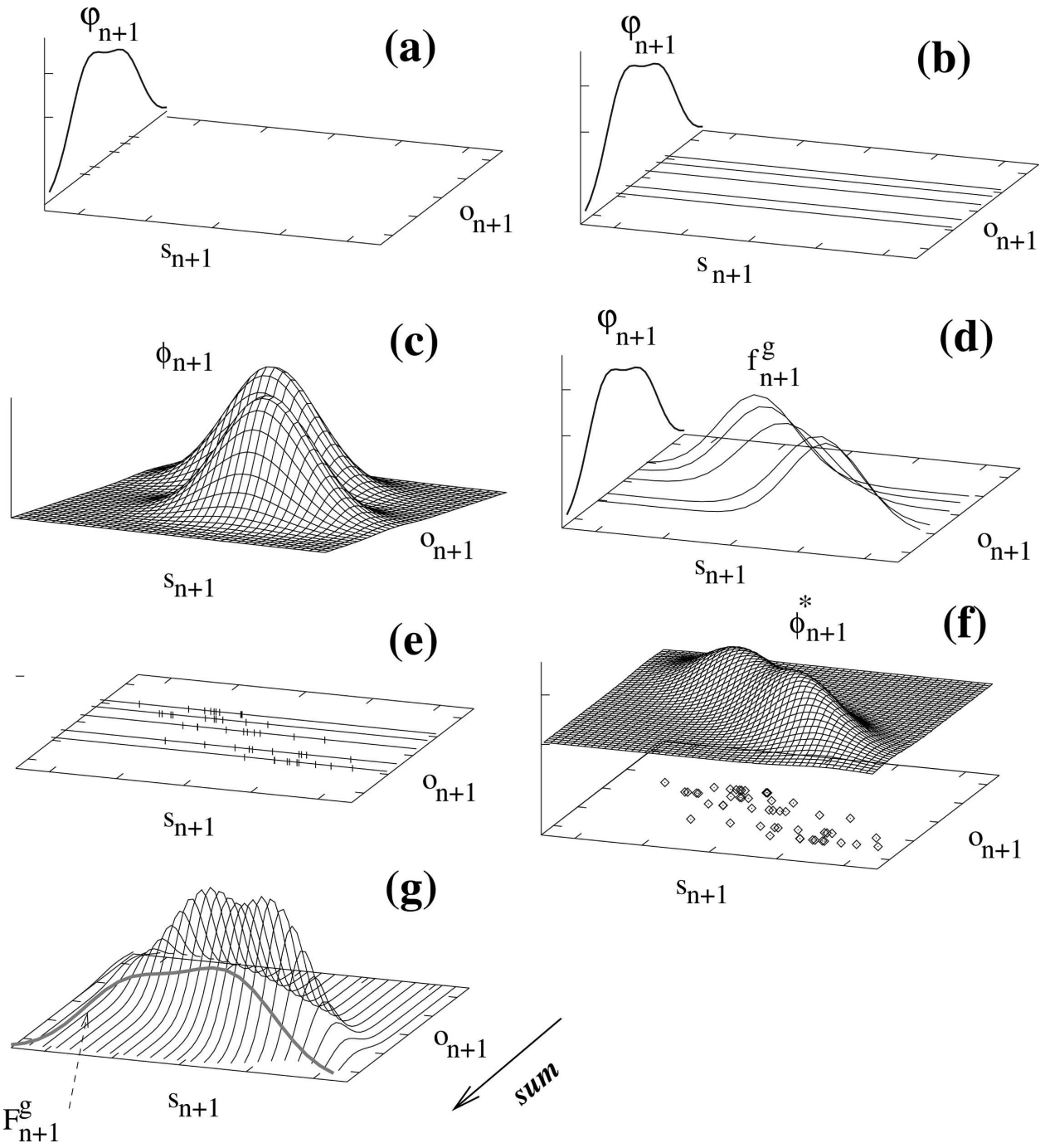

FIG. 2. Assimilation of observational pdf (φ_{n+1}) by GEnKF: Monte Carlo estimation of the analysis step.

other words, instead of using high-dimensional joint pdfs representative of the entire spatial domain of a particular model, we propose to use a few lower-dimensional marginal pdfs that are representative only of subdomains. For marginal LE pdfs, identification of these subdomains might be based on two discriminatory variables described in section 7c.

5. Control run and the surrogates

When estimating MG pdfs for bivariate LE ensembles, there are three potential difficulties that should be addressed:

- *Undersampling:* For optimization of 2D MGs, six parameters have to estimated for each component [see

Eq. (1)]. The algorithm in Figueiredo and Jain (2002) is only guaranteed to be robust with regard to the initialization procedure if the sample size is "large enough." As mentioned in section 4 the number of points required for reasonable density estimation in D-dimensional space scales roughly as M^D. So gaps in data might result in the higher uncertainty in MG parameters. There are a few common reasons for missing records in satellite-derived data. For example, the performance of the RS retrieval algorithms for surface temperature T_s is influenced by cloud cover. Thus, SEBS predictions are available only for cloud-free or partly cloud-free days. Moreover, satellites provide only a snapshot view of spatial variability in T_s and as a consequence indirectly the spatial variability in LE_{sat}. Finally, accidental technical failures in in situ or RS instruments are responsible for gaps in data.

- *Instrumental and model errors*: Although LE_{is} observations are the closest approximation to natural variation of LE at local scales, the techniques used to measure LE_{is}, as, for example, EBBR systems, are themselves not without error. Indeed, in the heterogeneous landscape of the First International Satellite Land Surface Climatology Project (ISLSCP) Field Experiment (FIFE) campaign, LE_{is} predictions were often as high as 20% in error (Nie et al. 1992). On the other hand, LE_{sat} estimates are prone to errors in RS inputs to SEBS and limitations of SEBS itself to reproduce the complicated physical situation in the surface layer of air.
- *Scaling*: The footprint over which an LE_{sat} is determined is rarely at the same scale as LE_{is}, making direct comparison difficult.

The above-mentioned problems are usually responsible for strong scattering effects that blur the dependency structure underlying the bivariate LE ensembles. Recalling that the second objective of this work is to investigate the robustness of a practical methodology for regionalization of marginal LE pdfs—or, in other words, to investigate whether for classes of visually similar regions particular pdfs can be representative—such a situation needs to be resolved. As a pragmatic approach we propose the following course of action. We first pose the hypothesis that $p(LE_{is}, LE_{sat})$'s have a structure that depends on land use and vegetation cover, equivalent to visually similar regions. Next, the hypothesis is verified by creating "idealized" bivariate LE samples for a variety of environmental conditions (in terms of water supply, available energy, saturation deficit, turbulent transport, and vegetation characteristics). These hourly, daylight-based samples are referred to as the *control run*. The LE_{is} in the control run are taken "as is," and LE_{sat} proxies are obtained from SEBS forced with R_n estimated from in situ–measured radiation fluxes and T_s derived from the longwave radiation:

$$T_s = [(LW_{out} - (1 - \epsilon)LW_{in})/(\epsilon\varsigma)]^{0.25}, \qquad (10)$$

where LW_{out} denotes the outgoing longwave radiation, LW_{in} refers to the incoming longwave radiation, ϵ is the emissivity of the surface, and ς stands for the Stefan—Boltzmann constant. Note that in SEBS R_n and H, which specify the total available energy and sensible heat flux, respectively, depend on T_s. Next, the LE_{sat} in the control run is perturbed by Monte Carlo propagation of "satellite" error in T_s through R_n and H in SEBS (see section 7a for technical details of this procedure). This way, keeping again LE_{is} unchanged, we obtain another bivariate sample referred to as the *surrogate data*. It is expected that introduction of error sources blurs the structure in the control run but does not let it disappear. Thus, MG pdfs fitted separately to control run and surrogate data should be similar. To quantify the strength of this similarity we use the L^2 correlation in Scott and Szewczyk (2001):

$$C_{L^2}(p_1; p_2) = \frac{\int_{-\infty}^{+\infty} dx\, p_1(x)p_2(x)}{\left[\int_{+\infty}^{-\infty} dx\, p_1^2(x) \int_{+\infty}^{-\infty} dx\, p_2^2(x)\right]^{0.5}}, \qquad (11)$$

where p_1 and p_2 are L^2 integrable pdfs for the control run and surrogate data, respectively. This measure is 0 if two pdfs show no similarity and 1 if two pdfs are just the same. For MGs (11) can be calculated analytically.

To summarize the above procedure: undersampling in LE_{sat} data is tackled by creating an hourly surrogate data; by creating a control run of hourly data, error propagation is controlled and scaling is accounted for automatically by obtaining the conditional pdfs $p(LE_{is}|LE_{sat})$ in (2) from regionalized marginal pdfs $p(LE_{is}, LE_{sat})$ and using these conditionals to compute the regression curves in (3).

6. Data

The measurements of LE_{is} used in this study come from six EBBR U.S. Department of Energy's Atmospheric Radiation Measurement Program Cloud and Radiation Testbed (ARM/CART) stations (E15, E4, E9, E20, E7, E25) distributed across the SGP region of the United States (see Fig. 3). ARM is aimed at obtain-

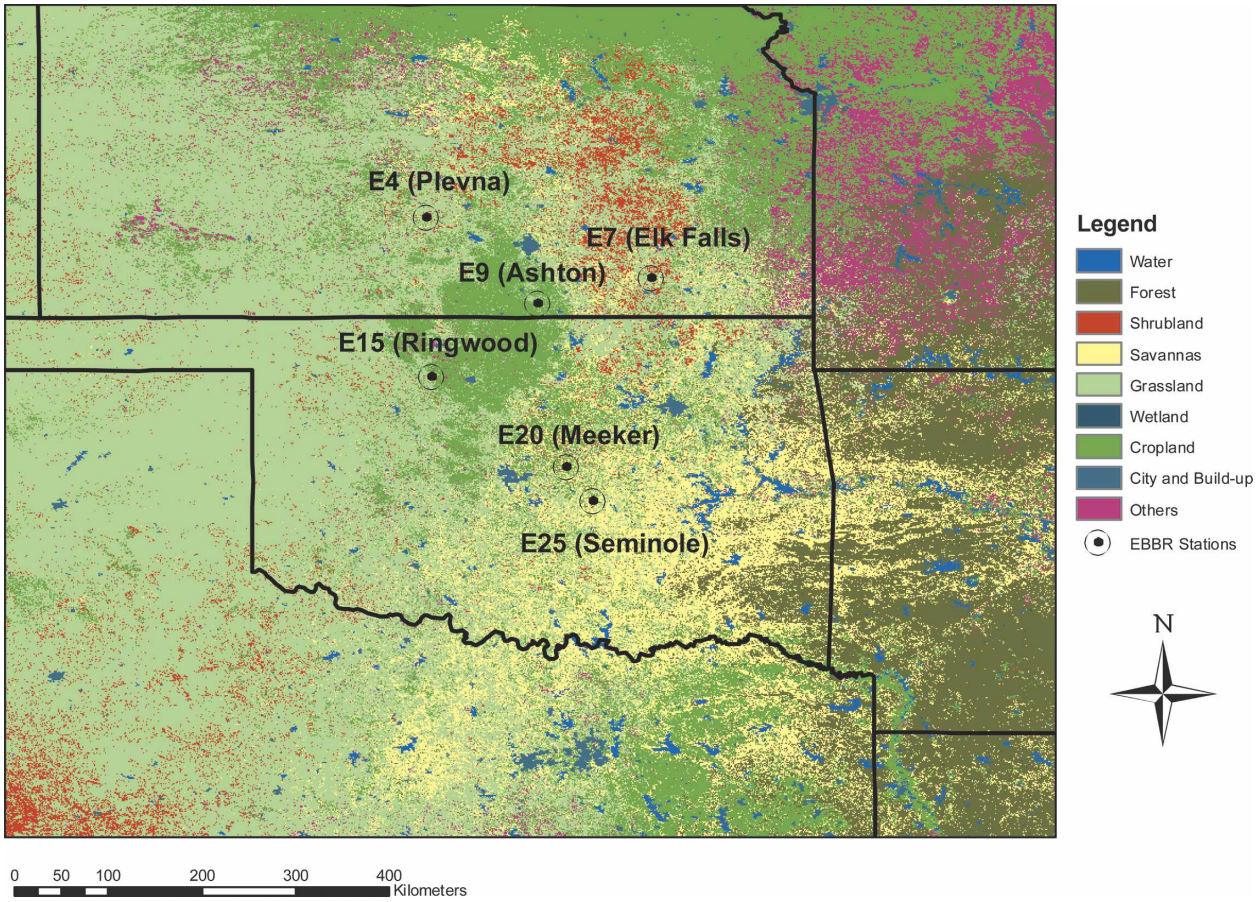

FIG. 3. Land-cover classification from MODIS in the International Geosphere–Biosphere Program (IGBP) scheme for the ARM/CART region in Oklahoma. Circles represent the distribution of EBBR stations across the region.

ing field measurements and developing models to better understand the processes that control solar and thermal infrared radiative transfer in the atmosphere and at the earth's surface. The SGP CART site was the first field site established by ARM and consists of in situ and remote sensing instrument clusters across north-central Oklahoma and south-central Kansas.

The LE_{is} proxies are based on 30-min averaged observations. The LE_{sat} estimates in the control run were obtained with SEBS forced with the input set displayed in Table 1.

Both types of LE data were obtained at 1-hourly resolution in the period of 1 July 2001–30 September 2001. We further restricted the data to 8-hourly sections of a day (0900–1700 local time) so we only considered unstable and neutral conditions in the atmospheric surface layer. Figure 4 shows the control run ensembles of the bivariate LE data.

These ensembles can be thought of being discrete samples from the unknown "true" marginal pdfs. Looking at the ensembles in Fig. 4 it is clear that simple parametric families of pdfs as Gaussians are not flexible enough to capture the geometry of the problem. For this reason in section 7b we fit MGs to describe the LE data.

7. Results

a. Analysis of surrogate LE data

To obtain the surrogate data, "satellite" errors in T_s were then propagated through SEBS for recalculations of R_n and H to obtain Monte Carlo simulations. We assumed that errors in T_s are additive and follow N(0, σ^2) distribution where $\sigma = \pm 1.5$ K. This estimate is derived from studies on MODIS T_s retrieval algorithms reported by Sobrino et al. (2003) and Wan et al. (2004). To perform the Monte Carlo propagation, 40 points were generated at random from the error distribution and added to or subtracted from a particular T_s measurement in the control run. This operation was independently repeated for all available T_s estimates from (10). Because we restricted ourselves to considering

TABLE 1. SEBS input for the control run.

SEBS variable (unit)	Source	Temporal resolution	Spatial resolution
Surface temperature (K)	Derived from ARM/CART longwave radiation data	1 h	Point
Emissivity (–)	MODIS	1 day	1 km
Surface pressure (Pa)	ARM/CART	1 h	Point
2-m air temperature (K)	ARM/CART	1 h	Point
Surface specific humidity (kg kg^{-1})	ARM/CART	1 h	Point
Surface wind speed (m s^{-1})	ARM/CART	1 h	Point
Surface albedo (–)	ARM/CART	1 h	Point
Shortwave radiation (W m^{-2})	ARM/CART	1 h	Point
Longwave radiation (W m^{-2})	ARM/CART	1 h	Point
LAI (–)	MODIS	7 day	1 km
Vegetation fraction (–)	MODIS	7 day	1 km
Land cover (–)	MODIS	1 yr	1 km

only unstable and neutral conditions in the atmospheric surface layer, whenever a realization of $T_s < T_a$ (T_a denotes 2-m air temperature) appeared (sporadically) it was replaced with a regenerated value, the regeneration process being repeated until $T_s \geq T_a$. All the T_s realizations were then propagated through SEBS to obtain the surrogate LE data. The surrogates are displayed in Fig. 5.

There is another subtle point pertinent to the above algorithm. In SEBS, an error in T_s directly contaminates R_{net} and H estimates (see Su 2002), and in consequence aggravates the errors in LE_{sat}. However, the question arises of whether the error in T_s is representative for the "satellite" error in R_{net}. To investigate that issue the following analysis was done. First, by comparing in situ–measured $R_{\text{net,is}}$ with SEBS-esti-

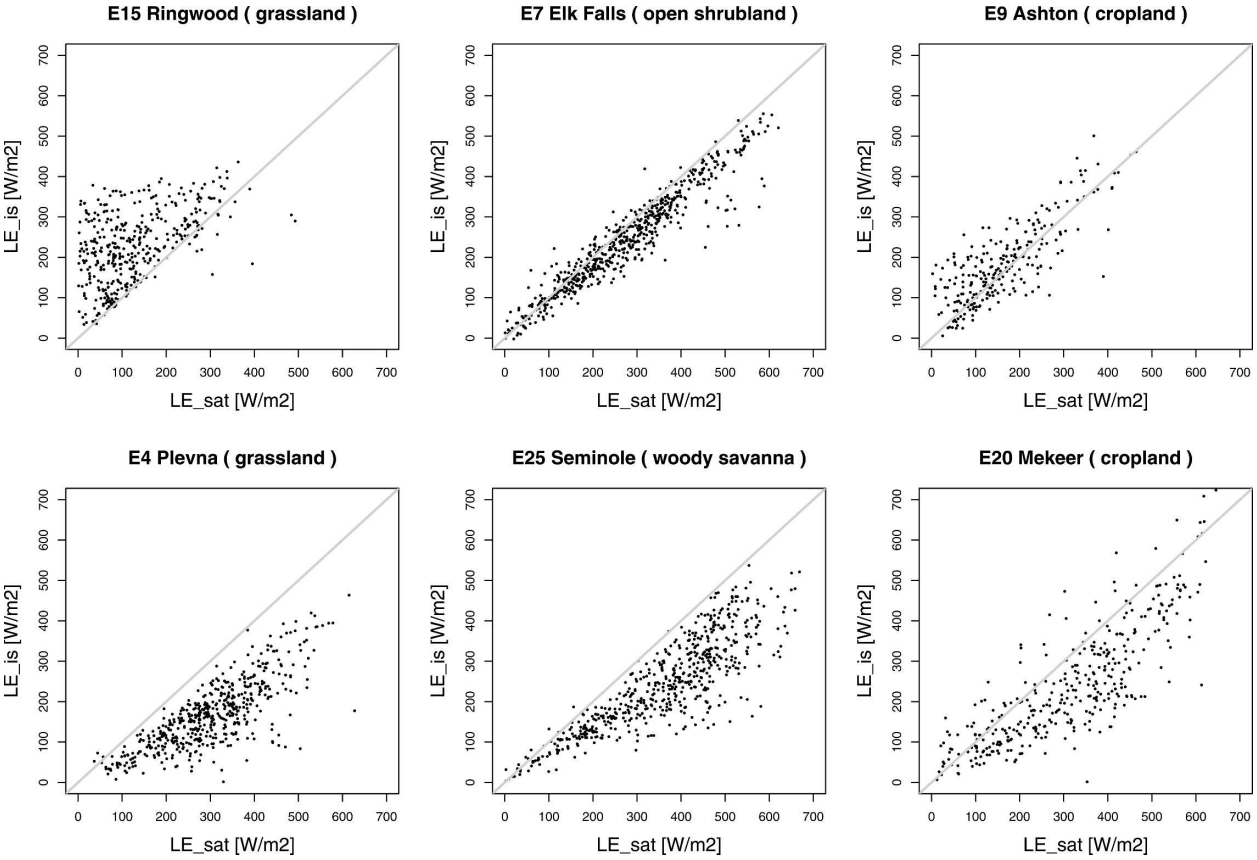

FIG. 4. The control run LE data for six sites in the SGP region for the period of 1 Jul–30 Sep 2001.

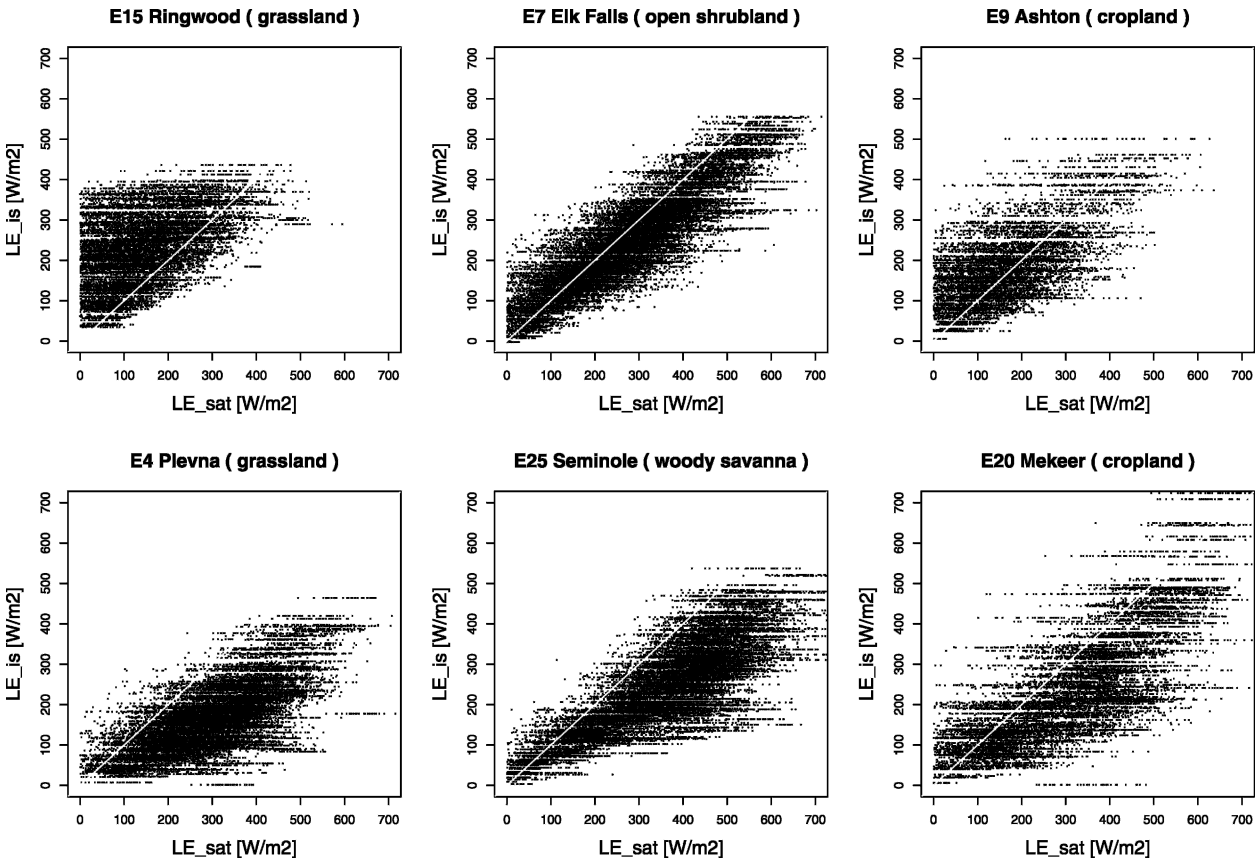

FIG. 5. The surrogate data for six sites in the SGP region for the period of 1 Jul 2001–30 Sep 2001.

mated $R_{\mathrm{net,sat}}$ using the dataset in Wood et al. (2003), we found the error in $R_{\mathrm{net,sat}}$ to be as high as a 15% coefficient of variation (cυ). Then, we performed a simple check on the order of magnitude of the absolute error in $R_{\mathrm{net,sat}}$ originating from the 15% error and from the 1.5-K error for a range of $R_{\mathrm{net,sat}}$ values. These results are demonstrated in Table 2.

The clear conclusion from the table is that the error in the T_s cannot account for the total error in $R_{\mathrm{net,sat}}$ as estimated by SEBS using the dataset in Wood et al. (2003). In practice there are evidently many more error sources in the computation of $R_{\mathrm{net,sat}}$, although we have to realize that part of the 15% cυ exists by the mismatch in spatial scale between $R_{\mathrm{net,sat}}$ and $R_{\mathrm{net,is}}$. Returning to the point of how serious H and R_{net} aggravate the contamination of $\mathrm{LE}_{\mathrm{sat}}$ by errors in T_s (cf. Figs. 4 and 5), it is mainly through H and only to a small extent through R_{net}. This also implies that we may expect in practice substantially more scatter in a figure like Fig. 5 if more error sources would be considered. However, here we restrict our exercise to the most simple case of a single error source in T_s to demonstrate the methodology and usefulness of applying MGs.

TABLE 2. Analysis of errors in $R_{\mathrm{net,sat}}$.

$R_{\mathrm{net,sat}}$	T_s	$R_{\mathrm{net,sat}} \pm 15\%$	$R_{net,sat} \pm 15K$
150 W m^{-2}	290 K	±22.5 W m^{-2}	±8 W m^{-2}
300 W m^{-2}	293 K	±45.0 W m^{-2}	±8 W m^{-2}
450 W m^{-2}	296 K	±67.5 W m^{-2}	±9 W m^{-2}
600 W m^{-2}	300 K	±90.0 W m^{-2}	±9 W m^{-2}
750 W m^{-2}	305 K	±112.5 W m^{-2}	±9 W m^{-2}

b. MG density fitting

Bivariate MGs were fitted to both the control run and surrogate data. We initialized mean $\mathbf{m}_n$ vectors in (1) to 30 randomly chosen data points. The initial covariances were made proportional to the identity matrix $\mathbf{C}_n = \sigma^2_{\mathrm{init}}\mathbf{I}$ with the diagonal entries σ^2_{init} equal to 1/10 of the mean of the variances along each dimension of the data:

$$\sigma^2_{\mathrm{init}} = \frac{1}{10D}\,\mathrm{trace}\left[\frac{1}{K}\sum_{i=1}^{K}(\mathbf{x}^{(i)} - \mathbf{m})(\mathbf{x}^{(i)} - \mathbf{m})^{\mathrm{T}}\right], \tag{12}$$

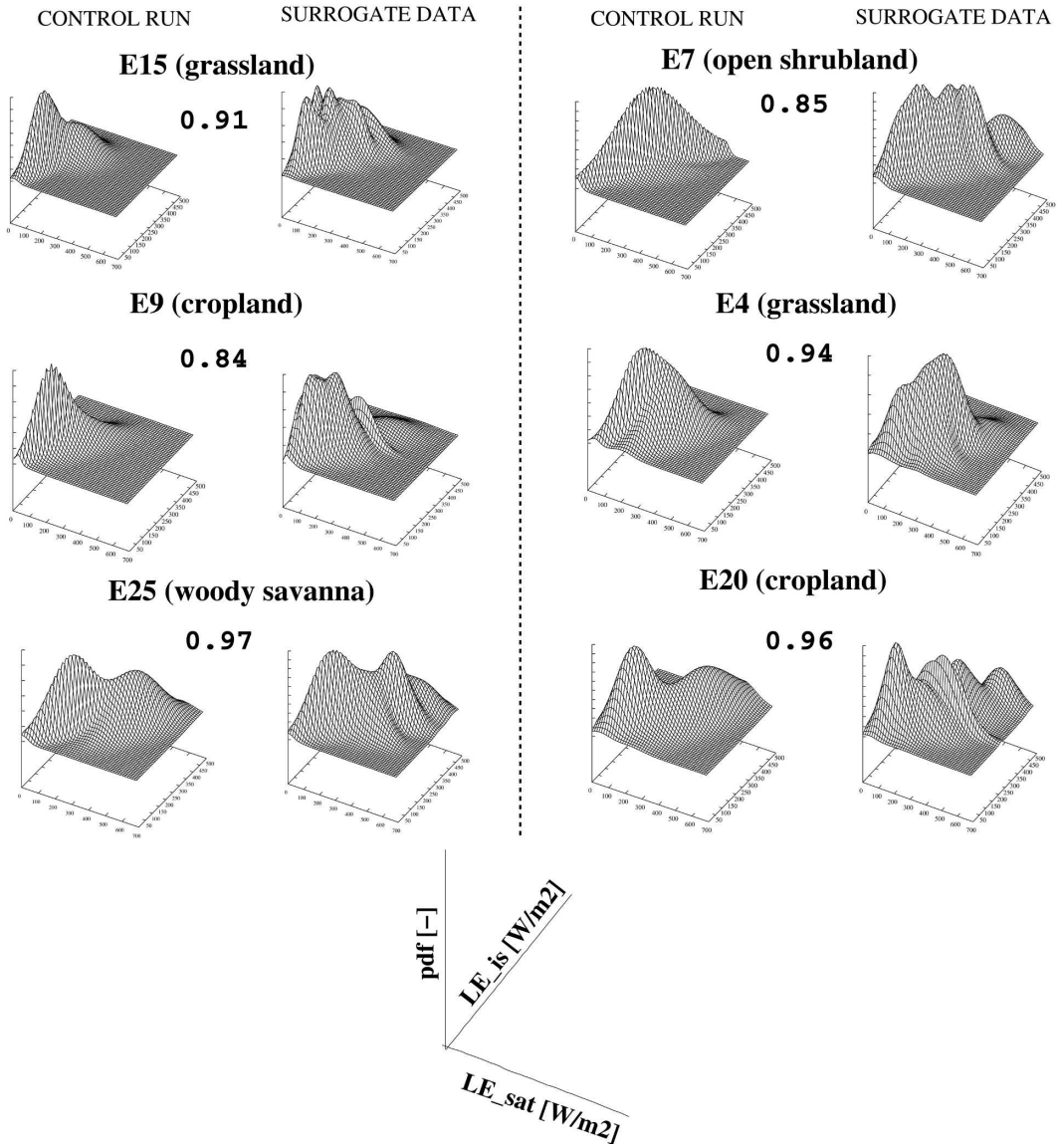

FIG. 6. MG pdfs $p(\mathrm{LE_{is}}, \mathrm{LE_{sat}})$ fitted to the control run and the surrogate data in Figs. 4 and 5, respectively. Numbers between corresponding pairs of pdfs indicate $C_{L^2}(p_1;p_2)$ in (11) estimated for these pdfs.

where $\mathbf{m} = (1/K)\Sigma_{i=1}^{K}\mathbf{x}^{(i)}$ is the global data mean (see Figueiredo and Jain 2002). This step was meant to assure the initial density on each data point be reasonably higher than 0. Figure 6 displays the fitted MGs. Comparing these pdfs to discrete underlying ensembles in Figs. 4 and 5 shows that MGs are a smoothed continuous representation of the underlying points. It is also clear from the figure that MGs can capture the particular local features of the ensemble while standard parametric densities (as, e.g., Gaussians) are unable to do this.

To quantify the similarity between pdfs for the control run and those for the surrogate data, the L^2 correlation in (11) was calculated for each pair of MGs. These results are given as numbers in Fig. 6. Since the value of the correlation is high (0.84–0.97), the error in T_s did not have much influence on the control run pdfs of the bivariate LE data. It refines the control run pdfs, one may say. Note that the error in T_s has, however, a pronounced impact on $\mathrm{LE_{sat}}$ estimates from SEBS, which can be seen by comparing horizontal spread of the ensembles in Figs. 4 and 5.

Calculation of the conditional MGs from the marginal MGs fitted to the surrogate data provides a basis for potential applications in spatiotemporal interpolation and data assimilation. The former can be

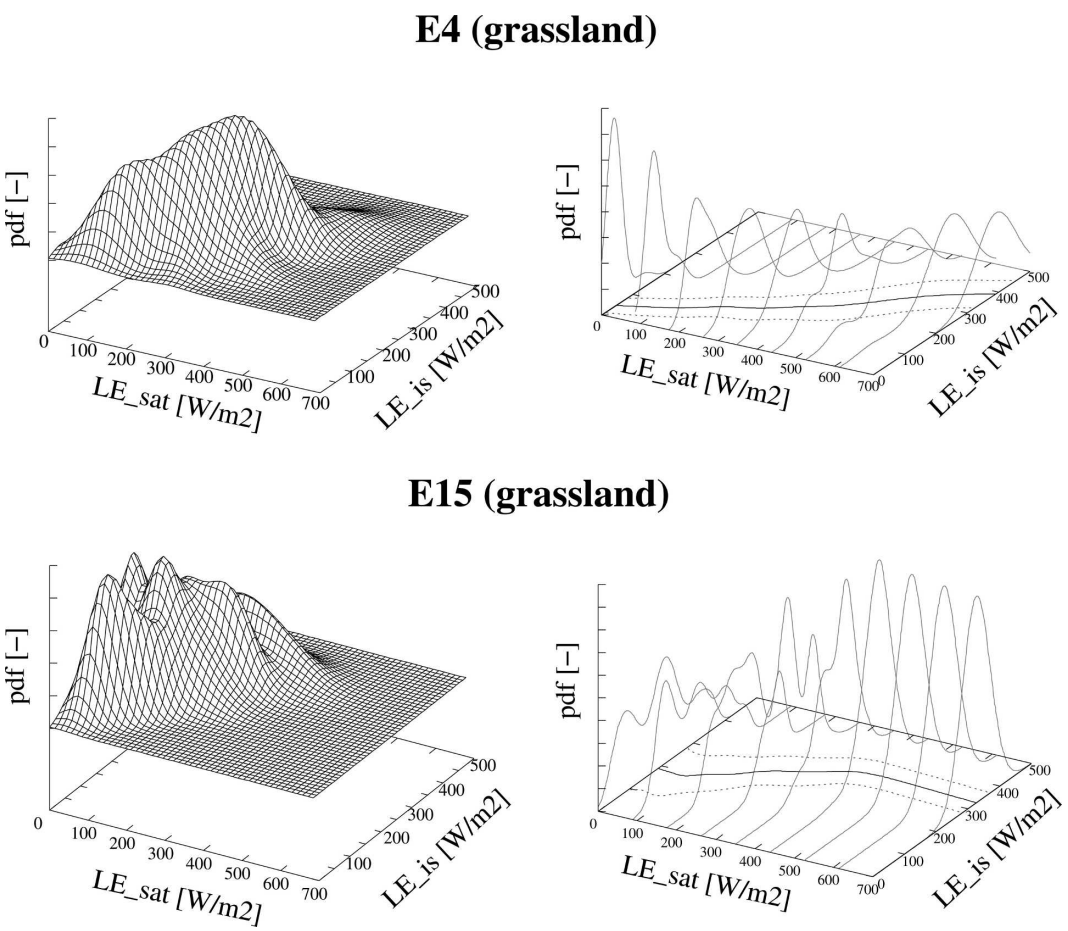

FIG. 7. (left) MG pdfs $p(\mathrm{LE_{is}}, \mathrm{LE_{sat}})$ fitted to the surrogate LE data for two grassland sites in the SGP region. (right) A few examples of conditional MG pdfs $p(\mathrm{LE_{is}}|\mathrm{LE_{sat}})$. The solid line in the x–y plane represents the conditional expectation (regression curve) whereas the dashed lines represent the standard deviation envelopes.

achieved by making probabilistic predictions of $\mathrm{LE_{is}}$ by either resampling from the conditional density $p(\mathrm{LE_{is}}|\mathrm{LE_{sat}})$ or calculating the conditional expectation $E[\mathrm{LE_{is}}|\mathrm{LE_{sat}}]$. The latter requires the knowledge of $p(\mathrm{LE_{is}}|\mathrm{LE_{sat}})$ and implementation of the algorithm described in section 4. For two grassland sites in the SGP region, marginal and conditional pdfs together with corresponding regression curves and standard deviation envelopes are presented in Fig. 7.

It is clear that the form of the conditional MG pdfs alters with an increase of $\mathrm{LE_{sat}}$ values and reveals a variety of shapes: from Gaussian to highly non-Gaussian (e.g., multimodal or skewed). This nonstationary behavior influences the geometry of conditional expectation and standard deviation envelopes, which, for E15, are clearly nonlinear. Interestingly, the conditional pdfs for the E4 site appear to flatten (or technically, to have higher entropy) with increasing value of $\mathrm{LE_{sat}}$. The opposite is true for the E15 site. This is in accordance with the scatter patterns for both sites in Fig. 4. Both sites are situated in the grassland area and show a similar range of $\mathrm{LE_{is}}$ values. However, it can be seen from Figs. 4 and 5 that hydrometeorological conditions, as produced by SEBS, for E15 suggest that it is dryer than E4. Therefore, our first conclusion with respect to the objectives of this study is that land-cover type alone cannot be used to regionalize LE pdfs. In the next section we identify an additional control parameter in SEBS that makes the regionalization feasible.

c. Regionalization of MGs

One of the SEBS parameters that characterize vegetation cover over a particular area is vegetation fraction (f_c). This parameter, which takes values from 0 to 1, plays a role in the estimation of soil heat flux and most importantly in the determination of scalar roughness height for heat transfer (Su 2002). The latter is a crucial parameter in parameterization of the momentum and heat transfer. From the physical point of view

one may say that the smaller the f_c, the earlier reduction of LE_{sat} with respect to some potential LE will start and reduction will follow a steeper slope for a period of dryness. In contrast after a period of dryness the LE of the $(1 - f_c)$ fraction will restore much more quickly than that of the f_c fraction.

To test the sensitivity of the bivariate LE data to f_c we performed the following exercise. We extracted the $<$ min; max $>$ range of f_c for E4 and E15 sites. This range was $<$ 0.45; 0.61 $>$ and $<$ 0.22; 0.32 $>$, respectively. Then for E4 we altered the original values of f_c by subtracting the [0.1; 0.2; 0.3; 0.4] offset and for E15 by adding the [0.2; 0.4; 0.5; 0.6] offset in the control run, respectively. Afterward, SEBS was forced with these eight variants of f_c while keeping the other variables in the control run unchanged. So, for each of the two sites we obtained four different bivariate LE ensembles. These ensembles are shown in Figs. 8 and 9.

Clearly, f_c controls the spread of the presented ensembles around the identity line. In the case of E15 ensembles one can notice an acceleration in the evapotranspiration process. The opposite effect is visible for E4. Notice that by adjusting the value of f_c, the initial geometric pattern of the E15 ensemble (see Fig. 4) can be approximately transformed into the pattern present in E4 ensemble (cf. lower-right panel of Fig. 9 with the lower-left panel of Fig. 4). We repeated the same exercise for the remaining sites. In all cases we were able to achieve the similar degree of control by shifting the ensembles depending on the initial geometry of the ensemble in the control run. So practically we may conclude that land-cover type together with land-cover intensity f_c are steering factors for regionalization of LE pdfs. From the remaining scatter, we do realize, however, that there are many other variables responsible for LE dynamics.

Ideally, regionalization of bivariate LE pdfs would require us to collect LE_{is} observations for a range of f_c within a land-cover type, which is difficult to achieve in practice, especially for sparse in situ networks. However, as demonstrated above we can control the shape of bivariate ensembles to a reasonable extent by changing the value of f_c in SEBS. So the approximate solution here could be as follows:

- at a given site (a) with a particular land-cover type and with available in situ LE measurements the bivariate pdf are fitted to data ensemble
- for a different site (b) with the same land-cover type and without the in situ latent heat flux measurements, estimated f_c (presumably from remote sensing) is used to transform the existing ensembles at site (a) refit the pdf

To demonstrate an example of the above approach the following cross-validation procedure was performed. Given information about bivariate LE fluxes at E4, Plevna site with grassland cover, and known f_c we tried to infer the structure of the MG pdf at E15, Ringwood site, with the same land cover but different f_c. So, basically, we fitted the MG densities to the data ensembles in Fig. 8 {i.e., we approximated f_c at the E15 site by f_c at E4 minus the offset [0.1; 0.2; 0.3; 0.4]} and compared these with the MG density for the E15 site fitted to the control run ensemble in the upper-left panel of Fig. 4. The comparison was done using the similarity measure in (11). The results are displayed as numbers (in bold) in Fig. 8. It is easy to see that the L^2 correlation between pdfs is highest (0.81) when subtracting the 0.4 offset, which yields a range of f_c between $<$ 0.05; 0.21 $>$. This range, however, is small compared to the original f_c range derived from MODIS for the E15 site: $<$ 0.22; 0.31 $>$. Therefore, there is some unexplained uncertainty in the regionalization procedure, which can be attributed to the uncertainty in MODIS-derived f_c and differences between the two sites in terms of the soil water balance. The latter source of uncertainty is demonstrated in Fig. 10 as a scatterplot between in situ evaporative fractions for E15 and E4. Clearly E4 is evaporating more than E15. This is related to available energy, soil moisture (thus antecedent precipitation), soil water storage capacity, and f_c (assuming vegetation types are the same). So one can conclude that by using land-cover type and f_c as a regionalization parameter, one gains a reasonable amount of information about underlying bivariate pdfs at sites where LE_{is} measurements are not available. This information, however, is not sufficient to guarantee full recovery of the underlying MG pdf, due to aforementioned sources of uncertainty. It is worth mentioning that similar cross-validation results were obtained for E9, Ashton, and E20, Mekeer, sites with cropland land cover.

8. Discussion and outlook

In this paper we have proposed a new procedure for describing bivariate LE ensembles as MG pdfs. This procedure is able to produce nonparametric pdfs that are fully data driven and are more flexible to describe local geometry of LE ensembles than classical parametric pdfs like, for example, Gaussians. Moreover, the procedure offers a vehicle for uncertainty analysis due to undersampling, error sources, and scaling problems, which are notorious when comparing LE_{is} to LE_{sat} data. We have shown that the conditional pdfs

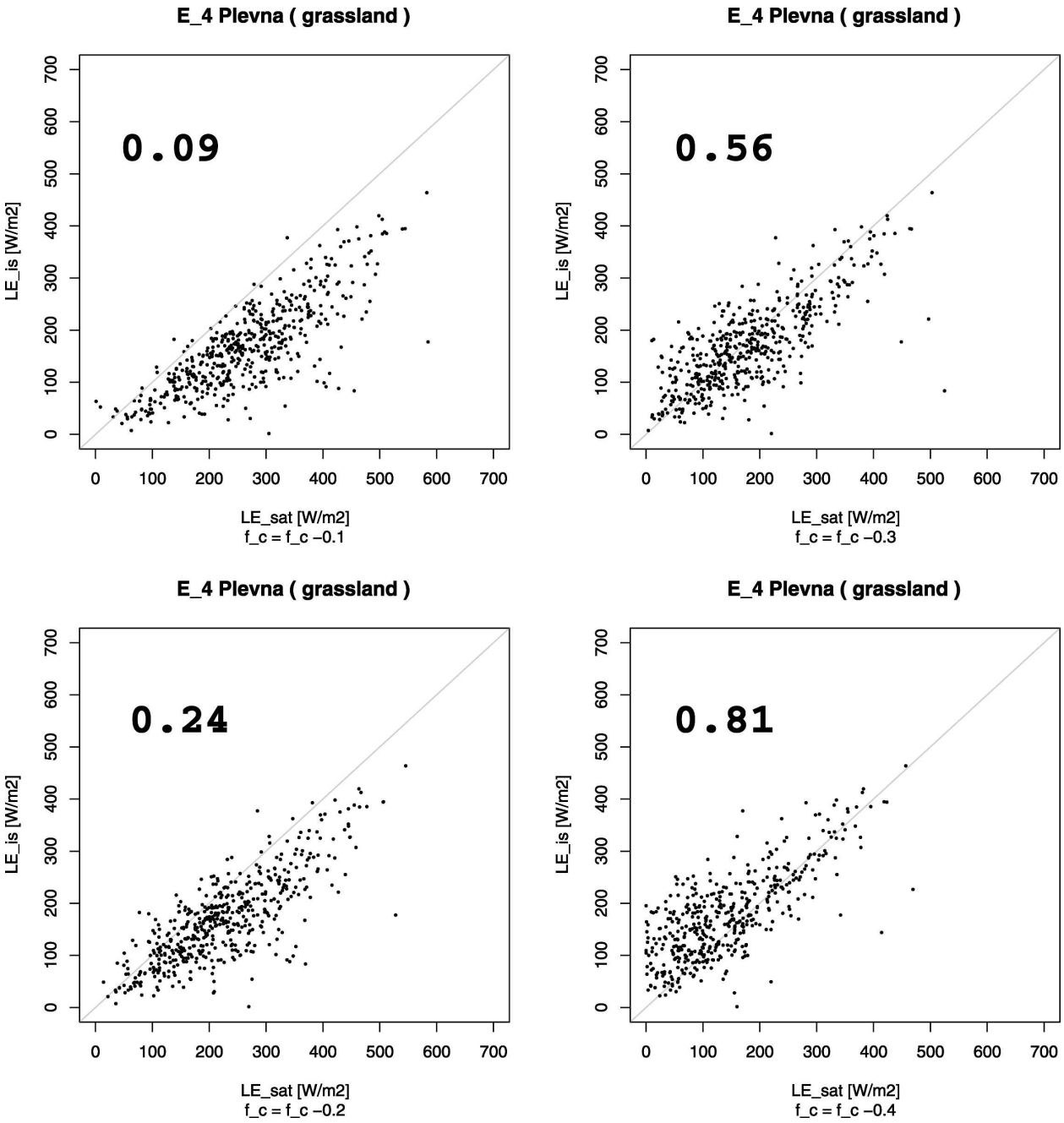

FIG. 8. Controlling the geometry of bivariate LE ensemble by subtracting an offset from f_c in the control run for the E4 site. Numbers displayed in upper-left corner of each scatterplot indicate $C_{L^2}(p_1;p_2)$ in (11) between the pdf fitted to each of the f_c altered ensembles and the pdf fitted to the control run ensemble for site E15 in the upper-leftmost panel of Fig. 4.

$p(\mathrm{LE_{is}}|\mathrm{LE_{sat}})$ can theoretically be useful in novel data assimilation schemes and spatiotemporal interpolation of bivariate LE data. The essential prerequisite for these applications is the ability to regionalize MG pdfs. The preliminary results in this work have demonstrated that it is feasible to regionalize the pdfs using land cover and vegetation fraction as discriminatory variables.

There are a number of issues that need to be investigated in order to implement the above methodology in hydrologic practice. Additional research is needed to quantify errors in $\mathrm{LE_{is}}$ data for various measuring techniques and see how they influence pdfs of bivariate LE ensembles. Progress in this direction, for EBBR/ARM-CART sites, is discussed in Stricker and Wójcik (2005,

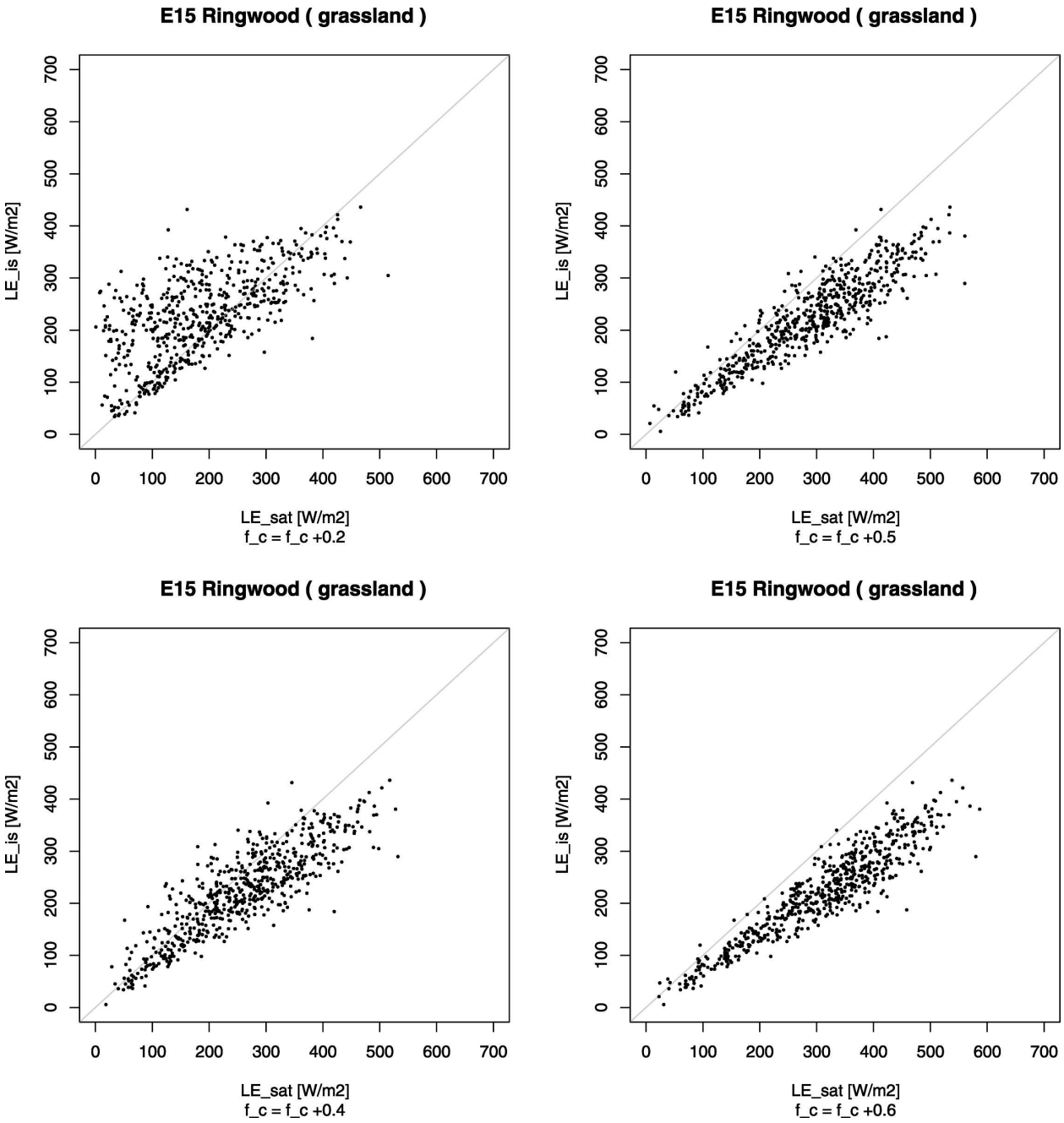

FIG. 9. Controlling the geometry of bivariate LE ensemble by adding an offset to f_c in the control run for the E15 site.

unpublished manuscript). An attractive option would also be to use scintillometeric measurements. The scintillation technique is one of the few techniques that can provide LE_{is} fluxes at scales of several kilometers (up to 10 km), making them more comparable to LE_{sat} fluxes (Meijninger et al. 2002). Additional work is further required to fine-tune the regionalization of LE pdfs by investigating the use of other environmental variables and influence of various uncertainty sources on pattern of bivariate LE ensembles. As demonstrated in section 7c an environmental variable to conceive here is the soil moisture. Finally, the parallel implementation of GEnKF for high-dimensional hydrologic systems is a challenging research problem. If successful, the assimilation of $p(LE_{is}|LE_{sat})$ into land surface models promises to enhance significantly the quality of water balance estimates over a range of spatial and temporal scales.

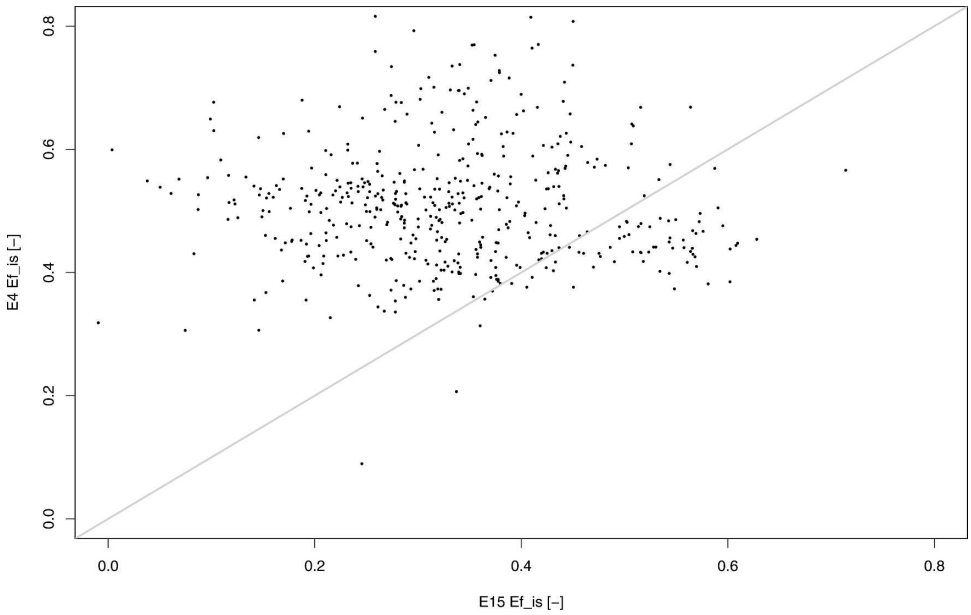

FIG. 10. Evaporative fraction at E15 vs evaporative fraction at E4 grassland sites for the period of 1 Jul–30 Sep 2001.

Acknowledgments. The first two authors are grateful for the financial support from WIMEK, the Wageningen Institute for Environmental and Climate studies. We also wish to acknowledge the constructive comments of Prof. S. Margulis and anonymous JHM reviewers whose inputs greatly improved the quality of our presentation.

REFERENCES

Anderson, J., and S. L. Anderson, 1999: A Monte Carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts. *Mon. Wea. Rev.,* **127,** 2741–2748.

Bellman, R., 1961: *Adaptive Control Processes: A Guided Tour.* Princeton University Press, 265 pp.

Bishop, M. C., 1995: *Neural Networks for Pattern Recognition.* Oxford University Press, 475 pp.

Braud, I., 1998: Spatial variability of surface properties and estimation of surface fluxes of a savannah. *Agric. For. Meteor.,* **89,** 15–44.

Brutsaert, W., 1999: Aspects of bulk atmospheric boundary layer similarity under free-convective conditions. *Geophys. Rev.,* **37,** 439–451.

Celeux, G., S. Chretien, F. Forbes, and A. Mkhadri, 2001: A componentwise EM algorithm for mixtures. *J. Comput. Graph. Stat.,* **10,** 699–712.

Evensen, G., 1994: Sequential data assimilation with non-linear quasi geostrophic model using monte-carlo methods to forecast error statistics. *J. Geophys. Res.,* **99,** 143–162.

Figueiredo, M., and A. Jain, 2002: Unsupervised learning of finite mixture models. *IEEE Trans. Pattern Anal. Mach. Intell.,* **24,** 381–395.

Gershenfeld, N., 1992: Dimension measurement on high-dimensional systems. *Physica D,* **55,** 135–154.

Hipps, L., and W. Kustas, 2000: Patterns and organisation in evaporation. *Spatial Patterns in Catchment Hydrology-Observations and Modelling,* R. Grayson and G. Bloshl, Eds., Cambridge University Press, 105–122.

Liang, X., D. P. Lettenmaier, and E. F. Wood, 1996: One-dimensional statistical dynamic representation of subgrid spatial variability of precipitation in the two-layer variable infiltration capacity model. *J. Geophys. Res.,* **101,** 21 403–21 422.

McLachlan, G., and T. Krishnan, 1997: *The EM Algorithm and Extensions.* Wiley Interscience, 352 pp.

——, and D. A. Peel, 2000: *Finite Mixture Models.* Wiley Interscience, 419 pp.

Meijninger, W., A. Green, O. Hartogensis, W. Kohsiek, J. Hoedjes, R. Zuurbier, and H. De Bruin, 2002: Determination of area-averaged water vapour fluxes with large aperture and radio wave scintillometers over a heterogeneous surface flevoland field experiment. *Bound.-Layer Meteor.,* **105,** 63–83.

Miller, D., J. Washburne, and E. Wood, 1995: Eos workshop on land-surface evaporation and transpiration. *Earth Obs.,* **7,** 52–56.

Nie, D., and Coauthors, 1992: An intercomparison of surface flux measurement systems used during FIFE 1987. *J. Geophys. Res.,* **97** (D17), 18 715–18 724.

Scott, D., and W. Szewczyk, 2001: From kernels to mixtures. *Technometrics,* **43,** 323–335.

Sharma, A., 2000: Seasonal to interannual rainfall probabilistic forecasts for improved water supply management: Part 3—A nonparametric probabilistic forecast model. *J. Hydrol.,* **239,** 249–258.

Sobrino, J., J. el Kharraz, and Z. Li, 2003: Surface temperature

and water vapour retrival from MODIS data. *Int. J. Remote Sens.,* **24,** 5161–5182.

Su, Z., 2002: The surface energy balance system (SEBS) for estimation of the turbulent heat fluxes. *Hydrol. Earth Syst. Sci.,* **6,** 85–99.

——, T. Schmugge, W. Kustas, and W. Massman, 2001: An evaluation of two models for estimation of the roughness height for heat transfer between the land surface and the atmosphere. *J. Appl. Meteor.,* **40,** 1933–1951.

Torfs, P., and R. Wójcik, 2001: Local probabilistic neural networks in hydrology. *Phys. Chem. Earth,* **26B,** 9–14.

——, E. van Loon, R. Wójcik, and P. Troch, 2002: Data assimilation by non-parametric local density estimation. *Computational Methods in Water Resources,* S. Hassanizadeh, R. Schotting, W. Gray, and G. Pinder, Eds., Elsevier, 1355–1362.

Wan, Z., Y. Zhang, Q. Zhang, and Z. Li, 2004: Quality assessment and validation of the MODIS global land surface temperature. *Int. J. Remote Sens.,* **25,** 261–274.

Wood, E., H. Su, M. McCabe, and Z. Su, 2003: Estimating evaporation from remote sensing. *Proc. IGARSS '03,* Vol. 2, Toulouse, France, IEEE, 1163–1165.