R. Wójcik, P.A. Troch, H. Stricker, P.J.J.F. Torfs, E.F. Wood, H. Su, and Z. Su (2004), **Mixtures of Gaussians for uncertainty description in latent heat flux estimates from remotely sensed information**, in *Proceedings of the 2nd international CAHMDA workshop on: The Terrestrial Water Cycle: Modelling and Data Assimilation Across Catchment Scales*, edited by A.J. Teuling, H. Leijnse, P.A. Troch, J. Sheffield and E.F. Wood, pp. 41–44, Princeton, NJ, October 25–27

Mixtures of Gaussians for uncertainty description in latent heat flux estimates from remotely sensed information

R. Wójcik[1], P.A. Troch[1], H. Stricker[1], P.J.J.F. Torfs[1], E.F. Wood[2], H. Su[3], and Z. Su[3]

[1]*Hydrology and Quantitative Water Management Group, Wageningen University, Wageningen, The Netherlands*
[2]*Princeton University, Princeton, New Jersey, USA*
[3]*Alterra Green World Research, Wageningen, The Netherlands*

Latent heat flux (LE) is the key variable that provides a link between energy and water budgets at the land surface. The conventional methods to estimate LE are based on point measurements of energy balance components and are representative only for very local scales. Recently a new class of techniques based on remotely sensed (RS) information has been developed to compute LE at scales from a point to a continent. Despite their potential, 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 the information about the true value of this quantity. The difficulty in inferring this information from data is due to different sources of uncertainty involved (e.g., measurement errors, scale problems, inadequacies in physical models that transform satellite observations into LE estimates). In this work we seek to investigate the use of non-parametric Gaussian mixture density models (GMDM's) to describe the conditional uncertainty of LE_{sat} given LE_{is}. This approach does not require any a priori assumptions on the form of the conditional density i.e. the algorithms we use in this study are completely data driven. An extra benefit from having the conditionals described by GMDM's is that they can further be applied to identify the recently developed non-linear Kalman filter for ensemble data assimilation (see *Anderson and Anderson*, 1999; *Torfs et al.*, 2002). This is the long run objective of this research.

Data and methods LE_{is} estimates used in this study come from seven Energy Balance Bowen Ratio (EBBR) ARM/CART stations (E15, E4, E9, E20, E7, E25, E8) distributed across the Southern Great Planes (SGP) region of the United States. These estimates are based on 30-min averaged observations. The LE_{sat} estimates were obtained using SEBS (Surface Energy Balance System) developed by *Su* (2002) and are based on instantaneous observations. The both types of LE proxies were obtained at 1 hourly resolution in the period of 1 July 2001–30 September 2001.

To describe the conditional uncertainty of LE_{is} **given** LE_{sat} a joint probability density function (pdf) f needs first to be fitted to bivariate sample $\{LE_{sat,k}; LE_{is,k}\}_{k=1}^{K}$. In this work the focus is on the use of GMDM's (see e.g. *McLachlan and Peel*, 2000) which are defined as linear combinations of Gaussian densities (see Figure 1.14), called components:

$$f(\mathbf{x}) = \sum_{n=1}^{N_c} w_n g_{(\mathbf{m}_n, \mathbf{C}_n)}(\mathbf{x}) \tag{1.1}$$

where $\mathbf{x}$ is a vector of variables, N_c the number of components, $g_{(\mathbf{m}_n, \mathbf{C}_n)}$ stands for the Gaussian density with mean $\mathbf{m}_n$ and covariance $\mathbf{C}_n$. Here $\mathbf{x} = [LE_{sat}\, LE_{is}]^T$. The w_n's are the component weights and satisfy $w_n \geq 0$ and $\sum w_n = 1$. Note that the conditional density $f(LE_{is}|LE_{sat})$ that is

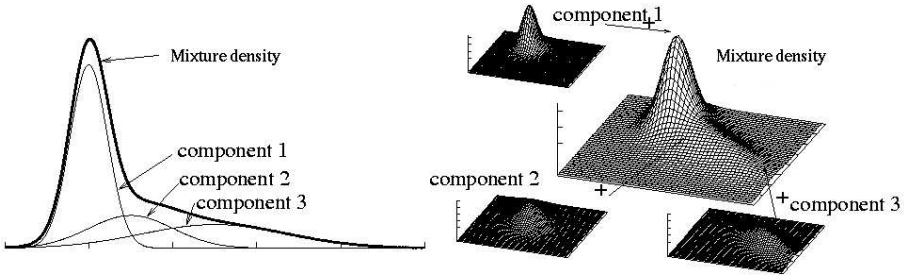

Figure 1.14: 1D and 2D example of GMDM (in both cases as a linear combination of 3 components).

calculated from (1.1) is also a GMDM. To fit (1.1) to the data the procedure of *Figueiredo and Jain* (2002) was applied.

First attempt To apply the above described methodology we derived the estimates of LE_{sat} by forcing SEBS with the following instantaneous RS inputs: short wave radiation derived from a 50 km GOES product and $1/8$ degree GOES surface temperature. The rest of the input variables that were needed to run SEBS (see *Su*, 2002) was either measured or taken from LDAS database. Next we grouped LE_{sat} and LE_{is} data according to landuse. Our hypothesis here is that at regional scale the bivariate dependency structure should be invariant within a particular landuse class. Moreover, this step is intended to tackle the dimensionality reduction issue in non-linear ensemble Kalman filters as described by *Anderson and Anderson* (1999) and *Torfs et al.* (2002). Figure 1.15 shows the result of this operation. It can be seen in the figure that the dependency *pattern* between LE_{is} and LE_{sat} is not really visible. Thus, the data in Figure 1.15 would be of little use for data assimilation purpose. The blurring effect might be due to undersampling which stems from the fact that data availability of GOES temperature is greatly affected by the cloud cover and the algorithm that is used to retrieve the surface temperature. Moreover, there is a spatial and temporal scaling problem involved (we compare point values with 1/8 decimal degree values), there is a measurement error in LE_{is} values and there is an error in LE_{sat} values. The latter might be a combination of errors in RS inputs to SEBS and limitations of SEBS itself to reflect the complicated physical situation in the near-surface layer of air. In what follows we address this issue by performing Monte-Carlo sensitivity analysis of SEBS to two RS inputs that in our view greatly influence the quality of LE_{sat} estimates: net radiation and surface temperature.

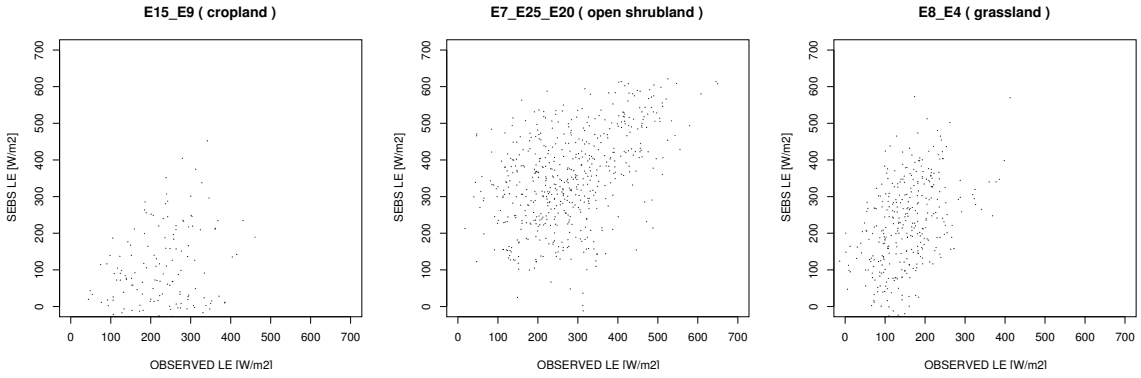

Figure 1.15: Bivariate 1-hourly *LE* data grouped according to landuse for the period 1 July 2001–30 September 2001. LE_{sat} estimates are derived from GOES products.

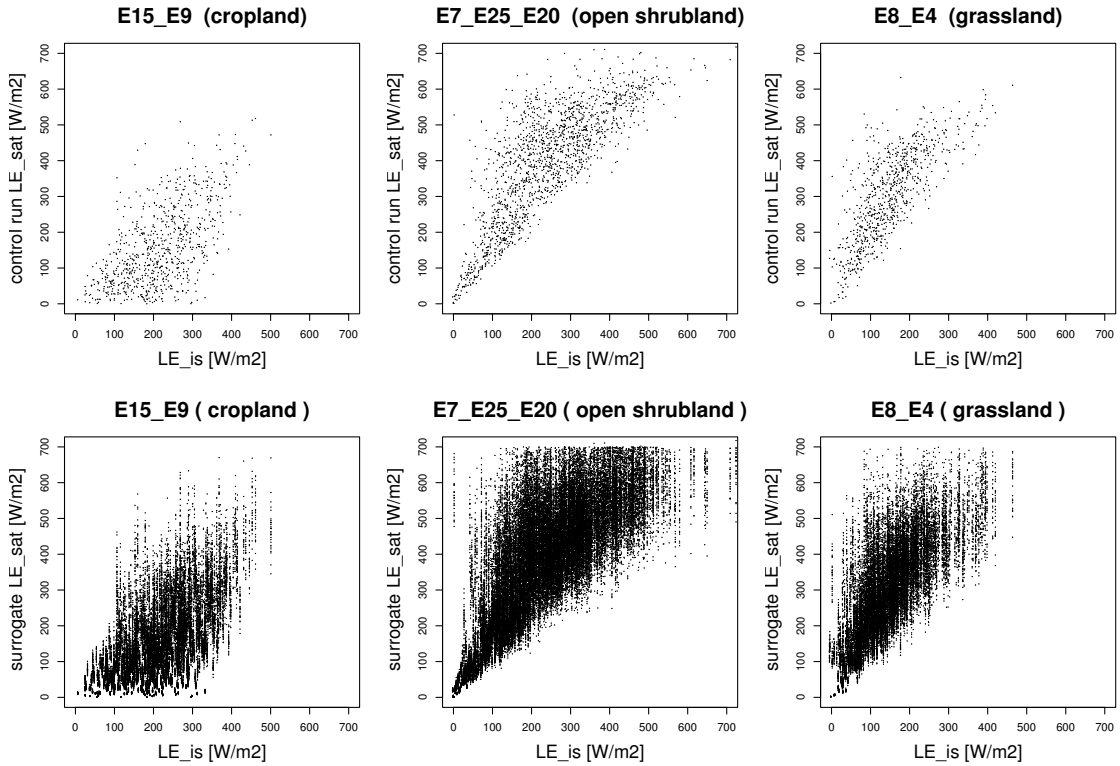

Figure 1.16: Bivariate 1-hourly *LE* data grouped according to landuse for the period 1 July 2001–30 September 2001. Upper panel: control run. Lower panel: surrogates.

Control run and surrogate RS data Accordingly, SEBS was forced with net radiation calculated from measured radiation components. Surface temperature was derived from outgoing long wave radiation. The rest of the inputs remained the same as mentioned in the previous section. In this way we obtained somewhat idealized LE_{sat} data (read: no RS error involved) which is reffered to as the control run (see upper panel of Figure 1.16). Note the transparent non-Gaussian dependency structure of bivariate *LE* data. Next, surrogate LE_{sat} data was created by perturbing the control run with percentual error in the net radiation (by comparing RS derived net radiation with measured net radiation we estimated this error as 15%). Technically, each net radiation measurement in the control run was treated as a mode of log-normal distribution and the 15% error as its coefficient of variation. From each distribution 30 points were drawn at random and propagated through SEBS to obtain LE_{sat} surrogates. Those are shown in lower panel of Figure 1.16.

Then, bivariate GMDM's were fitted to both control run and surrogate data from Figure 1.16 (for an example of fitted pdf's see upper panel of Figure 1.17). To determine to which extent the bivariate structure in control run was deteriorated due to sattelite error in net radiation we compared the fitted pdf's in terms of probabilistic similarity measure introduced by *Scott and Szewczyk* (2001):

$$sim(f_1; f_2) = \frac{\int f_1(x) f_2(x) dx}{(\int f_1(x)^2 dx \int f_2(x)^2 dx)^{\frac{1}{2}}} \tag{1.2}$$

This measure is 0 if two pdf's show no similarity and 1 if two pdf's are just the same. For cropland, open shrubland and grassland $sim(f_1; f_2)$ was 0.96, 0.96 and 0.98 respectively. This implies that

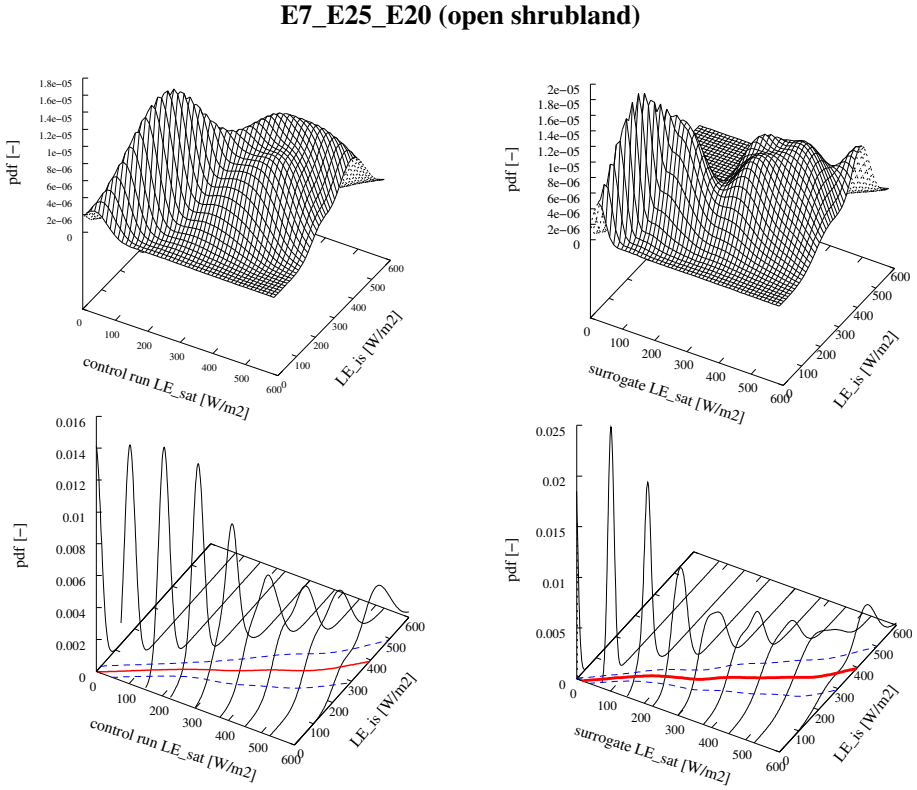

Figure 1.17: An example of MDGM's fitted to open shrubland *LE* data. Upper panel: joint pdf's $f(LE_{is}, LE_{sat})$ for control run and surrogates respectively (the similarity between the two pdf's is 0.96). Lower panel: conditional pdf's $f(LE_{is}|LE_{sat})$ for control run and surrogates. The solid line in X-Y plane represents conditional expectation and dashed lines represent standard deviation bands.

the error in net radiation has negligible effect on probability structure in control run data.

Continuation The same analysis will be performed for the surface temperature. The results will be shown during the poster session. In parallel we work on an uncertainty analysis of *LE* measurements from EBBR ARM/CART stations.

Bibliography

Anderson, J., and S. Anderson, A Monte Carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts., *Mon. Weather Rev.*, *128*, 1971–1981, 1999.

Figueiredo, M., and A. Jain, Unsupervised learning of finite mixture models, *IEEE T. Pattern Anal.*, *24*(3), 381–396, 2002.

McLachlan, G., and D. Peel, *Finite Mixture Models*, Wiley Iterscience, New York, 2000.

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

Su, Z., The surface energy balance system (SEBS) for estimation of turbulent heat fluxes, *Hydrol. Earth Syst. Sci.*, *6*(1), 85–99, 2002.

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