Effect of One-Dimensional Field Data Assimilation on Land Surface Model Flux Estimates with Implications for Improved Numerical Weather Prediction

Pipunic, R.C.[1], J.P. Walker[1], C. Trudinger[2] and A.W. Western[1]

[1] Department of Civil and Environmental Engineering, The University of Melbourne, Victoria, Australia
[2] CSIRO Marine and Atmospheric Research, Aspendale, Victoria, Australia
Email: r.pipunic@civenv.unimelb.edu.au

Keywords: *Data assimilation, latent heat flux, sensible heat flux, soil moisture, land surface model.*

EXTENDED ABSTRACT

The forecast quality from Numerical Weather Prediction (NWP) models and climate models depends on accurate initialisation. Therefore variables such as latent (LE) and sensible (H) heat flux from the land surface, which provide the lower boundary condition for NWP, need to be as accurate as possible at the beginning of a forecast period. Land Surface Models (LSMs) such as the CSIRO Biosphere Model (CBM) represent the exchange of energy and water between the earth's surface and lower atmosphere and are used to calculate LE and H. Soil moisture and temperature states of these models help partition incoming energy to the earth's surface between LE and H. Producing accurate predictions of LE and H is hindered by inaccuracies in LSMs such as uncertain initial model state conditions, errors in model forcing data, errors in model physics and a lack of data for accurately parameterising models.

Data assimilation blends observations of a model variable(s) with a model to update/correct the model and achieve more accurate predictions than by running the model offline. Assimilating soil moisture observations into LSMs is a proven technique for improving predictions of soil moisture and hence LE and H. Although, assimilating soil moisture may not necessarily lead to optimal LE and H predictions due to a complex and non-linear relationship between them. Assimilating LE and H observations has not been thoroughly explored in the scientific community and could potentially produce more accurate LE and H predictions. This study compares the assimilation of soil moisture observations with that of combined LE and H observations into the CBM with the resulting impacts on predictions of LE, H and root zone soil moisture and temperature examined.

Assimilation experiments were performed with a 1-year series of data using the Ensemble Kalman Filter (EnKF) algorithm. Observations and model forcing data were measured on a one-dimensional point scale at a site in south-eastern Australia. Errors were prescribed to initial conditions and to meteorological forcing variables. Observations were assimilated on typical remote sensing timescales – every 3 days for soil moisture (SMOS satellite) and twice daily with minimal cloud cover for LE and H (MODIS). Both observation sets were able to improve predicted soil moisture when assimilated compared to the offline model, with the soil moisture assimilation producing better results. Soil temperature predictions from both assimilation runs were worse than from the offline model indicating a warm bias. LE and H predictions are improved overall by both assimilation runs with LE and H assimilation producing the best predictions. It is demonstrated here that while surface soil moisture assimilation can improve soil moisture predictions in a LSM and consequently improve LE and H predictions, assimilating LE and H observations can produce more accurate LE and H predictions. Therefore the assimilation of LE and H observations into LSMs has the potential to provide NWP models with optimal LE and H estimates for initialisation.

1. INTRODUCTION

Latent (LE) and sensible (H) heat flux predictions at the earth's surface provide the lower boundary condition for Numerical Weather Prediction (NWP) and climate forecast models (Richter *et al.*, 2004), and hence are typically estimated by Land Surface Models (LSMs). Generating accurate NWP and climate forecasts requires getting the most accurate LE and H predictions possible from LSMs at the beginning of a forecast period. This can be achieved by adjusting the LSM soil moisture and temperature states to yield optimal heat flux estimates. Techniques such as data assimilation, which blend observed data with models to improve their predictive performance, are therefore useful in these cases. There are many examples in the scientific literature where data assimilation has been used to adjust LSM soil moisture states, on the basis that it will improve heat flux predictions for NWP. In particular, these published examples discuss assimilation of soil moisture (e.g. Walker and Houser, 2001) or screen-level (2m above ground) relative humidity and/or air temperature observations (e.g. Bouttier *et al*., 1993). However, assimilating observations of a variable such as soil moisture to correct LSM soil moisture states may not necessarily produce optimal heat flux predictions as the relationship between soil moisture and heat fluxes is non-linear and complex. This complex relationship is further exacerbated by the lack of availability in detailed soil and vegetation data required to parameterise LSMs. Achieving accurate and physically realistic soil moisture estimates should improve LE and H predictions but those predictions are not necessarily optimal.

A synthetic study by Pipunic et al. (2007-in press) has demonstrated that in addition to soil moisture assimilation, alternative approaches such as LE and/or H, and skin temperature assimilation also have strong impacts on improving LSM heat flux predictions. Moreover, it was shown that soil moisture and temperature predictions were also positively impacted by the assimilation. Therefore, this study extends that work to confirm the results when using real rather than synthetic data. Here, LE and H observations from a 3D eddy covariance system and soil moisture measurements are separately assimilated into the CSIRO Biosphere Model (CBM) (Wang et al., 2001) using an ensemble Kalman filter (EnKF) algorithm (Evensen, 1994), and the soil moisture, temperature and heat flux predictions compared with observed values. The aim is to make more definitive conclusions regarding which data type will likely lead to better LE and H predictions when assimilated into a LSM such as CBM.

2. STUDY SITE AND DATA

The modelling and assimilation in this paper uses data from a study site in south-eastern Australia, located within the Kyeamba Creek catchment. The site is approximately 30km south-east of the township of Wagga Wagga, New South Wales, situated on flat non-irrigated grass pasture land along the flats of Kyeamba Creek, a tributary of the Murrumbidgee River. Instrumentation at this site has been set-up and maintained by the University of Melbourne and includes a 3D eddy covariance system together with standard meteorological, soil heat flux (G), moisture and temperature profile measurements. Meteorological variables that were measured and used for forcing the CBM include incoming short and long wave radiation (outgoing short and long wave were also measured and used in determining net radiation, RN, for quality control of eddy covariance data), precipitation, air temperature, wind speed, specific humidity and atmospheric pressure. Any gaps in the meteorological record were filled with data from the Wagga Wagga automatic weather station operated by the Bureau of Meteorology.

Eddy covariance measurements were made at 10Hz and all other measurements at 0.5Hz with the exception of soil moisture (once every 0.00056Hz and atmospheric pressure (once per hour). All measurements were aggregated to 30 minute time steps for the experiment period January 1st to December 31st, 2005. The 3D eddy covariance system was elevated 3 metres above the ground giving an approximate fetch of 300 metres around the instrumentation. Data were quality controlled by i) filtering spurious values, ii) checking closure of the energy budget (scatter plot of $LE + H$ against $R_N - G$ revealed approximately 80% closure), and iii) closing the energy budget using the method of Twine et al. (2000), whereby the Bowen-ratio (H / LE) is maintained constant while LE and H are adjusted.

Key soil properties for parameterising the CBM in this study were sampled at the measurement site location using a map of soil data from the region (McKenzie pers. comms., 2005). They include wilting point, field capacity, hydraulic conductivity at saturation and bulk density. The values used in this study are assumed to represent the best possible parameters that would be available for use in the model. In contrast, most operational NWP models such as that used by Australian Bureau of Meteorology have globally uniform soil properties and Leaf Area Index (LAI) values (for vegetated areas) for parameterising the land surface (Richter et al., 2004). Vegetation

properties were assigned using estimates as in Sellers et al. (1996) for agricultural and C3-grassland, as supplied with the CBM. Parameter values that are observable in the field such as canopy height and fraction of roots in each of the model's soil layers were estimated at the site. LAI values were taken from monthly averaged LAI maps for Australia by Lu et al. (2001).

3. MODELS

3.1. CSIRO Biosphere Model (CBM)

The CBM as used in this study (the latest release version is called CABLE) was developed by scientists at the Marine and Atmospheric Research Division of the Commonwealth Scientific and Industry Research Organisation (CSIRO) in Aspendale, Victoria, Australia. A detailed description of the model and its formulations (written for the CABLE version) is given in Kowalczyk *et al*. (2006). It is a single column model dealing with the vertical exchange of water, energy and CO_2 between the soil, vegetation canopy and the atmosphere.

The soil scheme consists of six computational soil layers with thicknesses of 2.2, 5.8, 15.4, 40.9, 108.5 and 287.2 cm from top to bottom, all with uniform properties. There are three prognostic variables for each layer – soil moisture, soil temperature and ice content. Movement of water through the soil is governed by Richard's equation and calculation of heat conduction is used for determining soil temperature. Soil evaporation is modelled with a bulk aerodynamic formulation after Mahfouf and Noilhan (1991). Vegetation is represented by a two-leaf canopy model (Wang and Leuning, 1998) consisting of a big 'sunlit' and big 'shaded' leaf. Formulations are included for radiative coupling between the vegetation and ground, canopy turbulence, along with calculations of photosynthesis, stomatal conductance, leaf temperature, and energy and CO_2 fluxes. Total LE and H output from the model are the respective sums of LE and H from the soil surface and canopy.

3.2. Ensemble Kalman Filter (EnKF)

The EnKF is one type of direct observer assimilation methods. It can be summarised as follows:

$$\mathbf{X}_k^a = \mathbf{X}_k^f + \mathbf{K}(\mathbf{Z}_k - \mathbf{Z}_k^f), \qquad (1)$$

such that the state vector **X** forecast by the model (superscript f) at time k is updated (analysed; superscript a) by the difference between an observed and model predicted observation **Z** (the innovation) multiplied by a weight factor **K**. The weight factor, or Kalman Gain, is given by:

$$\mathbf{K} = \mathbf{P}_k^f \mathbf{H}^\mathbf{T} (\mathbf{H}\mathbf{P}_k^f \mathbf{H}^\mathbf{T} + \mathbf{R}_k)^{-1}. \qquad (2)$$

P represents the error covariance of the forecast model states and **R** is the error covariance of the observation. The matrix **H** is a nonlinear operator relating the state vector **X** to the observation **Z** (superscript **T** denotes the matrix transpose). If **P** is larger than **R** (i.e. observations more certain than model prediction), then **K** will approximate to 1 when **X** and **Z** are the same scalar quantity (i.e. **H** = 1), and the innovation will be relied upon heavily to adjust the forecast states due to the small relative observation error. In contrast, where **R** is large compared to **P**, **K** will approach 0 and the observation will not be trusted sufficiently leaving the final analysis vector $\mathbf{X}_k^a$ relatively unchanged, since the model's forecast is understood to be more reliable in this case.

A good summary of the EnKF as implemented in this study can be found in Walker and Houser (2005). The error covariance of the model, **P**, at assimilation times is estimated from a series of parallel model runs (ensemble members) each run with randomly added errors and **P** is calculated from the ensemble spread. The mean of the ensemble is therefore taken to be the estimate of the true state. At assimilation time steps, observations are perturbed within the observation uncertainty range and then an ensemble of observations is generated around the perturbed value within the uncertainty range, the spread of which is the observation error covariance, **R**. The model state matrix **X** is then updated via equations (1) and (2).

4. METHODOLOGY

This work compares observed LE, H and root zone soil moisture and temperature from the study site for all of 2005 with outputs from the CBM resulting from i) a model simulation without assimilation (denoted as "open-loop"), ii) a model simulation where LE and H observations are assimilated (denoted as "LEH_Assim") and iii) a model simulation where surface soil moisture observations are assimilated (denoted as "SM_Assim"). Initial conditions used in the model simulations were estimated by spinning up the model through repeated simulation using meteorological forcing data for the 1-year experiment period until the soil moisture and temperature states reached equilibrium at the

start/end of the year, which took 7 years of simulation. The measured forcing data and spun up initial conditions were used to produce the open-loop simulation. To implement the EnKF, ensembles of initial conditions and forcing variables were generated to represent modelling errors introduced by each – 20 ensemble members were found to be adequate. Inaccurate model physics and uncertain parameters also account for model error, but have not been specifically treated in this study. State updates were made to the soil moisture and temperature states for all six CBM soil layers in both assimilation experiments.

4.1. Initial condition ensembles

Generating initial condition ensembles involved adding random perturbations to the spun-up initial soil moisture and temperature values (taken to be the ensemble means) for each model soil layer. The perturbations were randomly generated variates with zero mean and a standard deviation chosen so the spread of ensemble members represents the uncertainty of the estimated initial conditions. Observed soil moisture and temperature at the initial time step were used to approximate the uncertainties of the estimated initial conditions. Since no observations were available to guide ensemble initial spread for the deeper model layers, the range of random values generated was made larger than for the other layers to reflect greater uncertainty. Table 1 summarises the spun-up initial conditions (ensemble mean) for each model soil layer and the range of ensemble spread (standard deviation) determined with the aid of observations.

Table 1. Summary of initial condition ensembles for soil moisture and temperature states.

	Moisture (vol/vol)		**Temperature (˚C)**	
Model Soil Layer	**Mean**	**St. Dev.**	**Mean**	**St. Dev.**
1	0.095	0.10	22.0	15.0
2	0.07	0.15	32.5	15.0
3	0.47	0.37	27.2	15.0
4	0.47	0.37	24.9	15.0
5	0.47	0.40	22.2	25.0
6	0.45	0.40	18.4	25.0

4.2. Forcing ensembles

The same principle used to create initial condition ensembles was applied to generate meteorological forcing data ensembles – random perturbations with zero mean and a maximum range ensuring the ensemble spread represents the data uncertainty were added to each variable. However, the level of detail in prescribing error to each meteorological forcing variable was more complex and the approach outlined in Turner et al. (2007-in press) was applied in this study. For every variable, two types of error were prescribed to create the ensemble member, with i) separate random perturbations generated at each 30 minute model time step in the experiment period and added to the data value (measurement error), and ii) a single random perturbation generated once and applied at each time step in the period (calibration / representation error).

Prescribing error to forcing data also depends on the data type of each variable as categorised by Turner et al. (2007-in press), which defines variables as either unrestricted, semi-restricted or restricted. Unrestricted variables are measured on a scale with no maximum or minimum bounds and errors are considered independent at any point on the measurement scale and added directly to each data value. Semi-restricted data have a lower or upper bounding limit and errors are usually proportional to measurements and are added as a percentage of measured values, so if a particular variable has a minimum bound of zero then there will be no error if a value of zero is recorded (such as with precipitation). Restricted data are measured on a scale with an upper and lower bound. Added errors can either be independent of measurements and truncated at the boundaries of the measurement scale. Alternatively they can be generated via a variable approach where error is a function of the measurement and the maximum error is added at the mid point of the measurement domain and reduces linearly to zero at the domain boundaries. An example of a restricted data type is cloud cover fraction in the sky where at the end points of the measurement domain, the sky is either completely clear or completely covered (low/no uncertainty) and more uncertain in the middle.

The range of ensemble spread prescribed to each forcing variable in this study was chosen to represent physical reality, whereby it represents the error from using data from a single site to represent a region for instance. Table 2 summarises the forcing variables by their data type category from Turner et al. (2007-in press) with the approximate maximum uncertainty range

(standard deviation) prescribed for generating ensembles.

Table 2. Summary of meteorological forcing variable ensembles.

Forcing Variable	Category	Std. Dev.
Shortwave in	Semi-restricted	35%
Longwave in	Semi-restricted	32%
Precipitation	Semi-restricted	40%
Air Temperature	Unrestricted	3˚C
Wind Speed	Semi-restricted	120%
Specific Humidity	Restricted	0.003g/kg
Pressure	Semi-restricted	2.5%

4.3. Data Assimilation

NWP and climate models produce spatially distributed forecasts and therefore any operational data assimilation scheme would be best served by spatially distributed observations such as remotely sensed data. Although this study is for a single one dimensional soil column, the assimilation experiments were performed using the temporal scales of relevance to remotely sensed observations.

For assimilation, LE and H eddy covariance measurements were sampled from the observational record on a twice daily interval (10:00am and 14:00pm), which approximately corresponds to a MODIS satellite timescale for thermal infra-red (TIR) measurements from which LE and H estimates are derived. Since clouds can obscure remotely sensed TIR observations, further sub-sampling of the twice daily observations was performed using cloud cover data from Wagga Wagga weather station. Observations were discarded for cloud cover of more than 3 oktas; 3 oktas was chosen as a significant proportion of an entire image would be useful for assimilation provided the pixels under clear sky are relatively contiguous. Moreover, observational data gaps are introduced as a result of data quality control procedures and instrument mal-function. Volumetric soil moisture data measured over the 0-8cm layer were sampled at midday once every 3 days. This approximates both the sampling depth and expected temporal scale of soil moisture observations from the Soil Moisture and Ocean Salinity (SMOS) satellite which is soon to be launched (Kerr et al., 2001). As a result of the data sampling, a total of 219 LE and H observations and 112 soil moisture observations were used in the assimilation experiments.

Uncertainty in the LE and H measurements was estimated from energy budget closure calculation averaged over the entire experiment period, which was about $30 Wm^{-2}$. Moreover, the soil moisture observation uncertainty determined from calibration of field data was $\pm$4%vol/vol which corresponds to the expected accuracy from SMOS. Soil moisture observations were assimilated into a depth averaged combination of the top two soil layers of the CBM (2.2 and 5.8cm thick respectively) which have a combined depth that is equivalent to the observation depth.

5. RESULTS AND DISCUSSION

To examine the impact of the assimilation experiments, model predictions of LE, H, root zone soil moisture and root zone soil temperature are compared with field observations. Root zone values refer to an average (weighted by soil layer thicknesses) of values across the top 3 model soil layers which contain plant roots. Figure 1 is a comparison of root zone soil moisture. LEH_Assim and SM_Assim can each improve the model predicted moisture in the root zone when compared with the open-loop simulation, with SM_Assim yielding more accurate results as expected. However, LEH_Assim improvement in soil moisture mainly in the first half of the year is encouraging and shows that this observational data stream does contains soil moisture information.

Figure 2 shows an overall poorer performance in root zone soil temperature in both assimilation cases compared to the open-loop simulation, except for a brief period (~days 100-150) where LEH_Assim closely matches the observed temperature (as does the open-loop simulation). Comparing the two assimilation runs, LEH_Assim has produced a better estimate of soil temperature in cooler/wetter periods, with SM_Assim generally performing better at warmer drier extremes.

There is a significant improvement in LE and H for the first half of the year (Figures 3 and 4) for both assimilation experiments, with LEH_Assim generally performing better especially for LE predictions. Throughout the middle part of the year, the open-loop simulation matches the observations reasonably well for both LE and H and the assimilation runs have minimal impact, although SM_Assim slightly overestimates H in parts of this period. In the latter part of the year from approximately day 280 onwards, predictions

of H (Figure 4) from all model simulations match the observations fairly well overall. For LE predictions (Figure 3), LEH_Assim is clearly better in matching observations from approximately day 310 onwards, with SM_Assim having almost no impact compared to the open-loop simulation.

Poorer soil moisture prediction from LEH_Assim in the latter half of the year coincides with good predictions of LE and H from LEH_Assim. Conversely, good predictions of soil moisture from SM_Assim coincide with poor LE estimates from approximately day 310 onwards.

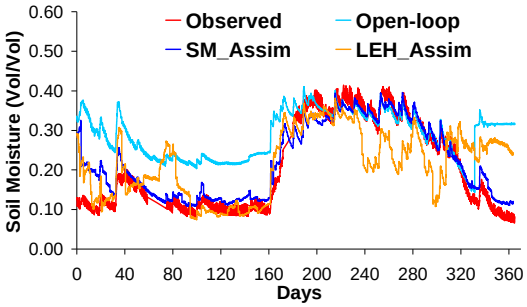

Figure 1. Root zone soil moisture observations and outputs from all model simulations.

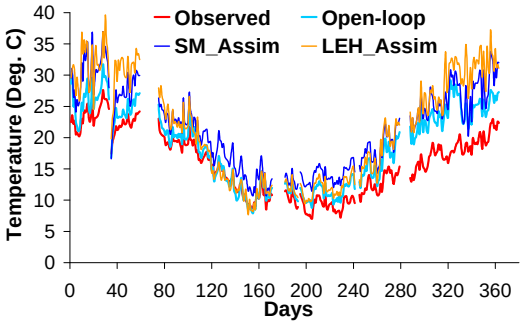

Figure 2. Daily averaged daytime (6am to 6pm) root zone soil temperature observations and outputs from all model simulations.

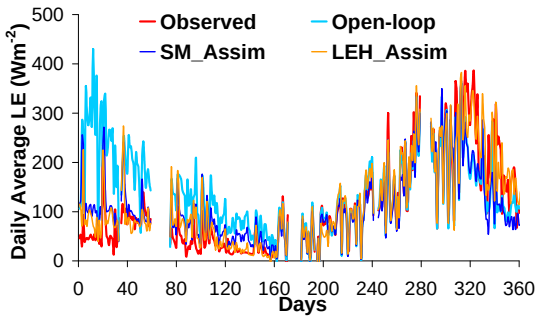

Figure 3. Daily averaged daytime (6am to 6pm) LE observations and outputs from all model simulations.

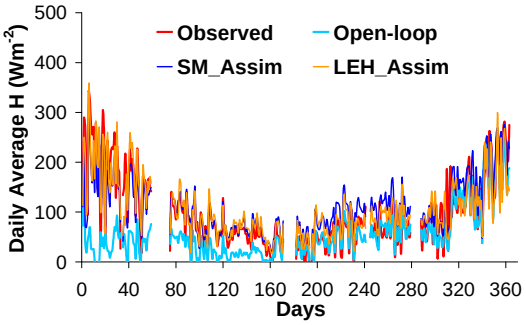

Figure 4. Daily averaged daytime (6am to 6pm) H observations and outputs from all model simulations.

Root mean square errors (RMSE) were calculated over the experiment period between model simulation outputs and the observations (Figure 5). They confirm improvements in all model outputs except for soil temperature from both assimilation runs. Also evident is the more accurate LE and H predictions from LEH_Assim compared to SM_Assim, and vice versa for soil moisture predictions.

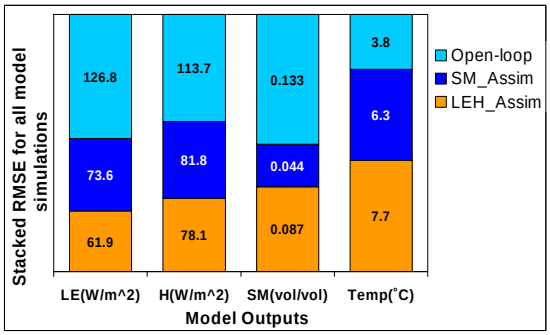

Figure 5. RMSE between model simulations and observations for daytime hours (6am to 6pm) over the 1-year experiment period.

6. CONCLUSIONS

The experiments performed here confirm: (i) SM_Assim has produced better soil moisture estimates than for LEH_Assim, with both producing better overall soil moisture predictions than the open-loop simulation; (ii) Both assimilation runs resulted in soil temperature warm biases, with LEH_Assim producing slightly better temperature predictions in the cooler/wetter period and SM_Assim better predictions in warmer/drier periods; and (iii) LEH_Assim produces better LE and H predictions than SM_Assim, with both performing better than the open-loop simulation.. Hence, improved soil moisture estimates do not necessarily translate to optimal LE and H estimates in LSMs and LEH_Assim has the potential to produce improved LE and H estimates for NWP.

7. REFERENCES

Bouttier, F., J.F. Mahfouf and J. Noilhan (1993), Sequential assimilation of soil moisture from atmospheric low-level parameters. Part I: Sensitivity and calibration studies, *Journal of Applied Meteorology*, 32, 1352-1364.

Evensen, G. (1994), Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics, *Journal of Geophysical Research - Oceans*, 99, 10,143-10,162.

Kerr, Y. H., P. Waldteufel, J-P. Wigneron, J-M. Martinuzzi, J. Font and M. Berger (2001), Soil Moisture Retrieval from Space: The Soil Moisture and Ocean Salinity (SMOS) Mission, *IEEE Transactions on Geoscience and Remote Sensing*, 39, 1729-1735.

Kowalczyk, E. A., Y.P. Wang, R.M. Law, H.L. Davies, J.L. McGregor and G. Abramowitz, (2006), The CSIRO Atmosphere Biosphere Land Exchange (CABLE) model for use in climate models and as an offline model. *CSIRO Marine and Atmospheric Research Technical Paper 013*, CSIRO.

Lu, H., M. R. Raupach and T. R. McVicar (2001), A robust model to separate remotely sensed vegetation indices into woody and non-woody cover and its large-scale application using AVHRR NDVI time series, *CSIRO technical report 35/01*, CSIRO.

Mahfouf, J. F., and J. Noilhan (1991), Comparative study of various formulations of evaporation from bare soil using in situ data, *Journal of Applied Meteorology*, 30, 1354-1365.

Pipunic, R.C., J.P. Walker and A. Western (2007), Assimilation of remotely sensed data for improved latent and sensible heat flux prediction: A comparative synthetic study, *Remote Sensing of Environment*, In press.

Richter, H, A.W. Western, and F.H.S. Chiew (2004), The effect of soil and vegetation parameters in the ECMWF land surface scheme, *Journal of Hydrometeorology*, 5, 1,131-11,46.

Sellers, P.J., S.O. Los, C.J. Tucker, C.O. Justice, D.A. Dazlich, G.J. Collatz and D.A. Randall (1996), A revised land surface parameterization (SiB2) for atmospheric GCMs. PartII: the generation of global fields of terrestrial biophysical parameters from satellite data, *Journal of Climate*, 9, 706-737.

Turner, M.R.J., J.P. Walker and P.R. Oke (2007), Ensemble member generation for sequential data assimilation, *Remote Sensing of Environment*, In press.

Twine, T.E., W.P. Kustas, J.M. Norman, D.R. Cook, P.R. Houser, T.P. Meyers, J.H. Prueger, P.J. Starks and M.L Wesely (2000), Correcting eddy-covariance flux underestimates over a grassland, *Agricultural and Forest Meteorology*, 103, 279-300.

Walker, J. P. and P. R. Houser (2001), A methodology for initializing soil moisture in a global climate model: Assimilation of near surface soil moisture observations, *Journal of Geophysical Research*, 106, 11,761-11,774.

Walker, J. P., and P. R. Houser (2005), In *Advances in Water Science Methodologies* (Ed, Aswathanarayana, A.) A. A. Balkema, Rotterdam, pp. 230.

Wang, Y.-P. and R. Leuning (1998), A two-leaf model for canopy conductance, photosynthesis and partitioning of available energy. I. Model description and comparison with a multilayered model, *Agricultural and Forest Meteorology*, 91, 89-111.

Wang, Y.-P., R. Leuning, H. A. Cleugh and P. A. Coppin (2001), Parameter estimation in surface exchange models using nonlinear inversion: how many parameters can we estimate and which measurements are most useful?, *Global Change Biology*, 7, 495-510.