Review of "Land surface parameter optimisation through data assimilation: the adJULES system" by NM Raoult, TE Jupp, PM Cox and CM Luke, submitted to CMD

Reviewed by Professor Hoshin Gupta (The University of Arizona, Tucson, AZ, USA)

Objective of this paper: To demonstrate application of automated local-search parameter optimization methodology to calibrate the parameters of the Joint UK Land Environment Simulator (JULES) land- surface model against eddy covariance measurements of gross primary production (GPP) and latent heat (LE) fluxes. The approach is used to define optimised parameter values (along with uncertainty estimates) for the 5 Plant Functional Types (PFTs) in JULES, improving the calibration and evaluation performance at 85% of the study sites used.

Data of GPP flux and LE flux from 160 sites were used. Input fields of vegetation structure and soil type were drawn from the UK Met Office files. LAI is derived from a MODIS product. The values taken for each site correspond to the closest grid point at which data are available.

Overall Comments: Overall I found the paper to be clear, well organized and well presented. As a report of the application of a specific gradient-based optimization method to parameter calibration of a specific LSM, the manuscript succeeds quite well. However, my assessment is that, as a general contribution to the scientific literature on the topic of LSM model performance improvement by application of data-assimilation for parameter estimation, the paper can certainly be improved.

For example the review of related literature [citing *Wang et al. (2001, 2007); Reichstein et al. (2003); Knorr and Kattge (2005); Raupach et al. (2005); Santaren et al. (2007); Thum et al. (2008); Williams et al. (2009); Peng et al. (2011), Xiao et al., (2011), Kuppel et al., (2012, 2014), Medvigy et al. (2009), Verbeeck et al. (2011), Medvigy and Moorcroft (2011), Groenendijk et al. (2010)*] fails to cite much of the related literature dating back to at least 1999 [e.g., *Gupta et al (JGR 1999), Houser et al (JGR 2001), Laplastrier et al (JGR 2002), Xia et al (JHM 2002), Demarty et al (JoH 2004), Liu et al (JGR 2004), Liu et al (JHM 2005), Demarty et al (WRR 2005), Hogue et al (WRR 2006), Abramowitz et al (JHM (2007), Rosolem et al (HP 2012)*, to name just a few].

While I certainly do not expect the authors to cite all of the above mentioned examples (which would be embarrassingly self-serving), I am generally concerned that the discussion does not show much awareness of the related developments in other closely related fields (such as Hydrologic Science) from which much of the impetus for optimization of the parameters of LSM's derives … beyond referring to the motivation being "*ideas from the applied mathematics of data assimilation as used widely in weather forecasting and other disciplines, and motivated by pioneering attempts at carbon cycle data assimilation (Rayner et al. (2005); Kaminski et al. (2013))*". Certainly, to my knowledge, much early work in LSM parameter optimization was promoted by the group led by Professor Pitman in Sydney. And as such, the general attention to parameter optimization considerably pre-dates the current interest in the broader concept of "*data assimilation*".

Further, I would generally expect a manuscript of this kind to provide a comparative evaluation of the *model performance improvements* with those of related parameter optimization studies performed by other members of the LSM community (albeit not using derivative based optimization). And while I am not arguing that a comparison with other kinds of *optimization methods* necessarily needs to be performed (although it would be useful and informative) I do

think it would be prudent to provide some comments about a) computational cost and b) relative advantages and disadvantages vis-a-vis other optimization approaches that have been used for LSM parameter optimization by the community (beyond simply remarking that the method is prone to premature convergence at local optima).

Nonetheless, I do have some more specific and serious concerns that are worthy of attention as indicated below. Please understand that my comments are intended to be helpful (from the perspective of my experience with such issues) to the authors to improve their manuscript and in no way should be interpreted as a criticism of the nice work that has already been performed and reported here.

Specific Comments:

Section on Methods and Data:

1) **Qn:** Only eight parameters relating predominantly to leaf-level stomatal conductance and photosynthesis (including the hydrological partitioning at the land-surface) are calibrated. Please comment on the fact that given there are *"over a hundred internal parameters" in JULES that need to be specified, and that "the detailed performance of a land-surface model can be very sensitive to such internal parameters",* it is quite possible that fixing most of the parameters during the optimization might affect the calibration results (due to parameter interdependence effects). Further, the model likely contains additional coefficients that are fixed (hard-coded) to values that may be generally suspect (see *Mendoza et al WRR 2015*); please comment on potential the implications of that to the results obtained.

   Mendoza PA, MP Clark, M Barlage, B Rajagopalan, L Samaniego, G Abramowitz and H Gupta (2015), *Are we unnecessarily constraining the agility of complex process-based models?* Water Resources Research

2) **Qn:** While you point out the inherent "*subjectivity*" and lack of reproducibility of LSM parameter calibration by manual adjustment, please comment on the history (since at least 1999) of the application of "*objective*" automated (multiple-criteria) methods to LSM calibration, albeit not with gradient based algorithms, and also please comment on the relative strengths and weakness of manual versus automated methods. I should also point out that the mathematical basis for such automated calibration significantly pre-dates the "*data assimilation*" literature that the authors cite, and goes back at least as early as Bard (1974, Nonlinear Parameter Estimation, Academic Press) as a well established reference.

3) **Qn:** Please comment on the reliability of the "*parameter uncertainty estimates*" provided by adjoint methods (linear-Gaussian approximation), given the significant parameter-output nonlinearity associated with LSMs (I note, in particular, your comment that "*optimal values need not be in the centre of the uncertainty range, the PDF can be skewed*". For example, does empirical Monte-Carlo sampling of the region of the optimum (thereby approximating the true non-Gaussian shape of the posterior parameter pdf) provide similar uncertainty ranges for the parameters?

4) **Qn:** Continuing on from the above, it appears that for the results you did not actually use the Hessian generated by adJULES to report the uncertainty estimates for the parameters, but instead used sampling of the posterior distribution (Section 2.5.1),

which makes perfect sense. Since the earlier part of the paper gives one the impression that the uncertainty estimates are computed directly by adJULES it would be good to modify the presentation to make the actual fact clear (remove possibility for confusion).

5) **Qn:** Please comment on the fact that use of an additive cost function (where only the total summed cost is minimized, as opposed to an approach where all individual cost functions are required to be improved) means that it is possible for the optimization method to achieve the "*best*" solution by improving the match at one site while possibly making the match worse (than it could otherwise be) at one or more other sites simply to achieve a better value of the cost function.

6) **Qn:** Please comment on the fact that, due to model structural errors, calibration to specific observables could actually cause model simulations of other fluxes (that were not used in tuning) to become worse (this is a potentially very serious problem in multi-flux calibration by weighted single criteria optimization to only some of the model fluxes); please see *Gupta et al (JGR 1999)*.

Gupta HV, L Bastidas, S Sorooshian, WJ Shuttleworth and ZL Yang (1999), Parameter Estimation of a Land Surface Scheme Using Multi-Criteria Methods, GCIP II Special Issue of the Journal of Geophysical Research-Atmospheres, Vol. 104, No. D16, p. 19491-19503

7) **Qn:** I am concerned that the ϵ_i^2 "*fractional variance explained*" measure is not really a properly informative measure of model performance, given that the benchmark for comparison is the observed seasonal cycle (I note also the related comment by a previous reviewer). In Hydrology this is related to a metric known as the "*Nash*" efficiency (equivalent to $1 - \epsilon_i^2$) that has been repeatedly demonstrated to be a poor index of model performance unless $1 - \epsilon_i^2 > 0.85 \text{ or } 0.9$ (such values are rarely achievable), and can hide the existence of significant bias in the performance (see, e.g., *Schaefli and Gupta (HP 2007)*). Note that the "Nash" efficiency is also typically justified as enabling cross-site and cross-model comparison, but arguable this is a poor reason for using a poorly informative metric. Instead the component decomposition [e.g., see *Murphy (Monthly Weather Review 1988)*, *Gupta et al (Journal of Hydrology 2009)*] can provide a more meaningful indicator of performance in terms of bias, variability and cross-correlation (see application in Rosolem et al, Hydrological Processes 2012).

Murphy A (1988), Skill Scores based on the Mean Square Error and their Relationships to the Correlation Coefficient, Monthly Weather Review 116, 2417-2424.

Rosolem R, HV Gupta, WJ Shuttleworth, LGG de Goncalves, and X Zeng (2012), Towards a Comprehensive Approach to Parameter Estimation in Land Surface Parameterization Schemes, Hydrological Processes, published online in Wiley Online Library (wileyonlinelibrary.com) DOI: 10.1002/hyp.9362

Gupta HV, H Kling, KK Yilmaz and GF Martinez-Baquero (2009), Decomposition of the Mean Squared Error & NSE Performance Criteria: Implications for Improving Hydrological Modelling, Journal of Hydrology, Vol. 377, pp. 80-91, doi: 10.1016/j.jhydrol.2009.08.003

Schaefli B and HV Gupta (2007), Do Nash values have value?, Hydrological Processes, 21(15), 2075-2080, simultaneously published online as Invited Commentary in Hydrologic Processes (HP Today), Wiley InterScience, doi: 10.1002/hyp.6825

Section on Results and discussion:

8) **Qn:** Please comment about your results in the context of the findings by *Abramowitz et al (JHM 2007)* who show that "*… as much as 45% of per-time-step model root-mean-*

square error in … flux outputs is due to systematic problems in … model processes insensitive to changes in vegetation parameters … <u>These results suggest that efforts to improve the representation of fundamental processes in land surface models, rather than parameter optimization, are the key to the development of land surface model ability</u>".

Abramowitz G, A Pitman, HV Gupta, E Kowalczyk and Y Wang (2007), Systematic Bias in Land Surface Models, Journal of Hydrometeorology, 8(5) pp 989-1001

9) **Qn:** Please comment on the sensitivity of the optimized results to choice of the constant of proportionality λ (how do the results change if λ is made smaller). Given the importance of using a "prior" on the parameters as constraint, this seems to me to be a rather important issue.

10) **Qn:** Please comment on the quality of the site-specific performance improvement in comparison with findings obtained by others. For example, *Rosolem et al (Hydrological Processes 2012)* reported that "*All sites showed improvements in simulation of the surface energy and carbon fluxes*" and "*In contrast, the default parameter sets (commonly used in GCM simulations) were found to be unable to reproduce the diurnal variation of energy fluxes at the tropical rainforest sites and showed a tendency to overestimate (underestimate) sensible (latent) heat fluxes. The calibration improved the simulations of these two fluxes by removing bias and variability errors (errors in signal mean and standard deviation)*."

11) **Qn:** In general it is well accepted that it is not ever possible to "*validate*" a model, and so the term validation in regards to model performance evaluation would seem to be misleading. Might I politely suggest the use of the more accurate term "*evaluation*" in its stead?

<u>Section on Conclusions:</u>

12) **Qn:** Perhaps it would be appropriate to comment on the computational cost involved with optimization using the adJULES system and on how many model runs (cost function evaluations) are necessary to achieve convergence (starting from the default parameter values)?