Review of the paper " Comparison of MODIS and SWAT Evapotranspiration over a Complex Terrain at Different Spatial Scales", by Olanrewaju O Abiodun, Huade Guan, Vincent E.A. Post, and Okke Batelaan.

The paper deals with the comparison of evapo-transpiration as estimated by the MOD16 satellite product and the SWAT hydrological model. The comparison is made on a 44 km2 semi-arid, mountainous basin in South Australia. The model is calibrated against observed discharged data, and the ET computed with the calibrated model is used for the comparison. The comparison is made on a range of spatial scales from 1 to 40 km2. The authors attempt to analyse the causes of the discrepancies between both products.

General comments

The papers addresses an important issue about estimating ET at the basin scale.
The paper is well written, with a clear structure. The language is fluent and precise. The methodology is generally well explained (see the detailed comments below). The paper is not totally novel as Et comparison studies are numerous, but it is probably novel regarding the region and the kind of basin used.

My main criticism is that the outcome of the paper remain, to me, a bit disappointing as the no clear conclusions are drawn on the cause of the biases between the two products (the discussion section mainly gives assumptions ore statements), nor any hierarchy between the possible causes. I made some suggestions below to help enrich the analyses. I agree, however, that such "timid" conclusions are inherent to this kind of analyses as the authors can not manipulate the satellite product to really test it, and as no reference ET product is available. May be adding other satellite product to the analysis (ex GLEAM) would be helpful ?

Besides, the authors do not give any informations on the usefulness of the study for other contexts or basins. The discussion may be more elaborated on this aspect,

Further, I do not see the fundamental motivation (obj # 2 of the paper) to compare both products on graduated spatial scales. I feel it is a way to evaluate the products rather than an objective as such.

The state-of-the-art section should be complemented with references on ET inter-comparison studies (models/satellite/numerical prediction systems), which could also enrich the discussion. See for example (I am not in the author's list !) Trambauer, et al, 2014, doi:10.5194/hess-18-193-2014. The authors should precise if SWAT has been used in such comparison studies. The novelty of the paper must be better emphasized.

As a conclusion, I think the subject of the paper is interesting but some complimentary analyses are needed before publication to reinforce the scope of the paper.

Detailed comments

l 33. the ref to Goyal et al seems not appropriate : this work is a bit old now (2004), and concerns a specific area on Inda. Better cite a/some refs dealing with a global perspective. Moreover I am not convinced that ET will be the most impacted everywhere. May be rainfall or even runoff might be severely affected. Please clarify.

l 50: …. *conducted a review* **OF** ? *30 remote....*

l 53-54 Moran and Jackson 1991 : add more recent reference(s)

Fig 1. Please give the meaning of all the variables (PET, Ecan, Et, Esoil, Revap, ET) in the caption to help understanding the figure.

l. 129 onwards (section 3.1). Some brief elements on the hydrology of the basin would be helpful. Please give an estimate of the mean annual discharge as compared to rainfall. Are the stream ephemeral (as suggested by fig 5), which simplifies the problem of setting the initial conditions for the simulations (see below) ? What is the depth to the ground water (see allusion on line 339). Is ground water the main source of river water ? What is the main sink for groundwater : ET or river discharge ? Impact of pumping ? …..

L 147 : the time range 2002 – 2016 is too short I think to be qualified as “historic”....

l 175 : do you mean : ”... all the 1 km 2 cells that **totally or partially** fall within the catchment area, i.e. you did not weighted the mean by the fraction of the cells overlapping the basin ?

l. 178 : How many HRUs did you use, and in what were their size range ? It could help understanding fig 6a.

L 177 onwards (Section 3.3).

It is unclear to me how you used, practically, the 5 objective functions for calibration, i.e. how you made the best compromise between all of them. Did you use the P- and R-factors metrics only to measure the confidence of the calibration once done, or also to select the parameter sets (as suggested at the end of line 216).

The way you calibrate the model is also unclear to me. It seems that you explored the parameter space (using Latin Hypercube Sampling) and selected a range of acceptable values for each parameters (see Table 3) based on the corresponding criteria values. Then which threshold values did you used for each criterion (obj. function ?)

By the way, the KGE criteria includes the three previous ones (r =R from eq. 4 ; omega = PBIAS from eq. 3 and alpha= Rsr from eq. 2), and is generally considered more robust than NSE, so I do not clearly understand why you used all 5 ?
Moreover lines 216-218 suggest that you finally used only P-factor, NSE, Rsr a,nd PBIAS to select the range of parameter values.

Please clarify all these points. May be a little more detailed description of the SUFI-2 algorithm will help understanding the procedure ? Please give a reference for this algorithm.

L 206-207. Referring to the paper by Gupta et al, (2009), *r* is the linear **correlation** (not regression) coefficient between the simulated and measured values

l. 213 : what does PPU stand for ?

L 220. Please detail how you managed the initial conditions at the beginning of each simulations period : spin-up period to equilibrate the internal reservoirs and mass budget of the model, or prescription (e.g. soil moisture, river discharges) from observations, or ... ?

l. 226-227 “*In the SUFI-2 algorithm an “r_” and a “v_” prefix before a SWAT model parameter indicate relative change and replacement change of the actual parameter values, respectively.*” : I do not understand what you mean here. Please clarify

l. 240 Table 3. As you come up with a range of value for each parameter after calibration, you get in fact an ensemble of hydrological simulations on each pixel or HRU's and on the whole period. I think the exploitation of this ensemble could be very fruitful, as it gives a kind of uncertainty range on your simulations (see for ex. Beven and Binley, 1992, *doi:10.1002/hyp.3360060305*). The comparison of MOD16 with the ensemble mean +- 1 standard dev, (for ex) would be informative.

l. 249, Table 4. Same comment than above (for section 3.3) which criterion were actually used for calibration, and what is the added value of R2, Pbias, Rsr as compared to NSE and KGE.

L 257, whole section 4.2.
I do not understand why you only evaluated the differences between the two products (figs 7 and 8), at various spatial scales, and not also (and firstly) their absolute value. It is difficult to clearly figure out which one is higher, where and when.
I suggest you display the empirical, statistical cumulated distributions (cdf) of the 1 km2 ET values for each product (for a given time aggregation period : 8-days, month, year, full period). They will inform on the relative magnitude of each series of value, the position of their means (I expect the mean of each distribution to be close to one another, as the basin-scale ET values are close). This analysis may reveal the causes of the biases. For example if both distribution are “close” or similar . it means that both product generate similar values as a whole but not necessarily over the same pixel or in at the same date (due to a bad land cover in MOD 12, and/or mismatch in the forcing fields...). It could explain, again, why the basin-scale means are close. I can also inform if this good match occurs for wrong reasons or only by chance ?.
May be the basin-scale averages converge mainly due to energy balance constraints, driven by the Penmann-Monteith equation, as at this spatial scale the atmospheric forcing for each model could be essentially comparable ?

An another informative analysis would be to compare MOD16 and SWAT ET on groups of pixels which correspond to comparable (or the most comparable) land cover on their respective map (cf fig. 4), even if their location do not match. (e.g. compare MOD16/Grassland and Tussock Grasses+ rainfed pasture ?). It could help check if a more realistic land cover (MOD12) would have produced a more realistic MOD16 ET.
I think this kind of analyses could enrich the discussion on the cause of the biases in section 5, which are, to me, a bit disappointing as nothing is really proven.

L 258-259 and Fig 6 b. It seems that you only considered the 1km2 MODIS cells fully enclosed within the basin boundaries. Please confirm.

L 276 - 208 and fig, 7. Please explain briefly how you aggregated to 2, 5, 10, .. km2 (Did you use the “spatial analyst tool in ArcGIS” as mentioned l 292-293 ?). Pixel grouping (2x2, 3x3, 4x4, ... pixels) results in areas of 4, 9, 16, … km2, which are not the aggregated areas you get. How did you manage the blank, 'no data' zones which inevitably fall into aggregated pixels ?

lines 278, 303, 428 “correlation” : except in the legend of fig. 9, it seems to me that you did not explicitly estimated the **correlation** between MOD16 and SWAT ET at each resolution. Please choose an other word or compute the correlations.

And by the way, I am not convinced that a lower max cell difference always indicates a better correlation (as suggested lines 277-279, unless I misunderstood) between the two series (here is a counter-example : consider a random series s(t), and two derived series s1=a.s(t) and s2=b.s(t), where a and b are scalars and a>b ; the correlation between s and s1 and s ans s2 is 1, but max(s1-s) > max(s2-s)). Please rephrase the section in a clearer way.

l. 282, fig 7. Does 41 km^2 correspond to the catchment scale ? It is said l. 140 that the basin area is 44 km^2. Please clarify.

l. 290, section 4.3.
What about comparing the dynamics at the lowest possible time step (that of MOD16 I guess). You rightly mentioned in the intro of the paper that despite its key role in the water cycle, ET was difficult to assess. Hence I think ET evaluation is also crucial at sub-monthly time steps.

l. 294 : R^2 and R are redundant, one of them is sufficient

l. 299, fig 9 (and figs 6, 7). Fig 6 visually suggests that MOD16ET > WAST ET at the catchment scale, which is supported by fig 7 (MOD16-SWAT >0), but which is not obvious from fig 9 where MOD16 ET seems < SWAT ET . Together with the differences MOD16-SWAT, it would be informative you give the absolute values of MOD16 and SWAT ET (in mm), e.g. at the catchment scale, on average on the whole period or year by year (see also my previous comment on that topic)

l. 343-345 : *"The convergence of the results of the two methods is also strongly attributed to the simple averaging …..from the MOD16 and SWAT ET to catchment scales."* Can you give more information to support or demonstrate this statement ? Which alternative averaging method(s) could be used ? Have you tested that they would impact the result at the catchment scale ?

L. 350 *"... deep rooted trees that can access the saturated zone... "* Is it the case on your basin ? Please give the information in the "study area" section.

l. 380-386. A realistic representation of rainfall intensities effectively improves streamflow simulations, specially in semi-arid areas, where Horton runoff dominates. However, but its impact on ET (which is mainly an inter-storm process) is probably weak, or weaker, all the more if you have calibrated the daily discharge, thus ensuring a consistent water balance in the basin. Please give more details on the links you see between high rainfall intensities and ET.

l. 426 *"....reliable ET estimates..."* please provide the basis you used to found this judgement. "Reliable" suggests you have a reference to compare with.....

l 430-431 *"...with two products derived from the MODIS satellite data classifying land cover differently..."* I do not understand : what are the 2 products derived from MODIS (MOD12 and ?)