Hydrol. Earth Syst. Sci. Discuss.,
doi:10.5194/hess-2017-147-SC1, 2017

Hydrology and Earth System Sciences
Discussions

Interactive comment on "Derived Optimal Linear Combination Evapotranspiration (DOLCE): a global gridded synthesis ET estimate" *by* Sanaa Hobeichi et al.

C. Jimenez

carlos.jimenez@obspm.fr

Received and published: 6 May 2017

I read with great interest this study about merging different terrestrial evaporation (ET) products. Given the current uncertainty of existing ET estimates, and that it is not yet demonstrated that a single methodology outperforms the others, avenues to merge products need to be explored, and I'm very glad to see efforts in that direction.

During my reading I took some notes, which I would like to share with the authors with the hope that they could contribute to the paper revision, or to future developments of DOLCE.

Product selection

While acknowledging the difficulties of finding the rights products to derive a synthesis product, I am a bit surprised about some of the choices. As the authors state, product selection should follow the criteria of product diversity, so ideally single algorithms with different strengths and weaknesses should be combined together. In that sense, I would have not considered an already synthesis product such as LandFlux-Eval as a possible candidate for the merge. In my opinion, a more valid alternative regarding LandFlux efforts could have been the three single products publicly available based on different ET algorithms (https://hydrology.kaust.edu.sa/Pages/GEWEX_Landflux.aspx). Also, I do not see much interest in combining obsolete versions of products with the current, and presumably better, product, as it has been done for GLEAM. Just GLEAM V3A (and perhaps GLEAM V3B) seems to me a better option.

An interesting product is MPI. This is a global extrapolation of the tower fluxes, the same tower fluxes that are used to decide on the merging weights, and quite different to the other products, which we could consider more "physically" based and less "calibrated" with the tower data. It is not surprise then that MPI is by a large margin the more weighted product. I do not deny that it can be a valid product for the merge, although much less independent than the other products with respect to the tower data. In that sense, it could have been very informative also to see how the merging works, and how the weights are distributed, when that product is left out.

Spatial and temporal resolution

The resolutions given in the Table 1 seem wrong for a number of products (e.g. MODIS original resolution is 1 km, Zhang 2010 is 0.05 deg, MPI and Zhang 2015 0.5 deg). Also, the periods of available data should also be revised for GLEAM (I think GLEAM V2A covered 1980-2011 and V3A 1980-2014).

As far as I can see, the only datasets limiting the study period to 2000 is MOD16 and GLEAM V2B. Perhaps the products going into merging could have been separated

in two groups, a bit similar to what has been done regarding geographical coverage: from a much earlier year than 2000 will all but MODIS and GLEAM V2B, and from 2000 including MODIS and GLEAM V2B. That would have resulted in a much longer DOLCE dataset, presumably based on a larger collection of tower data and more ET products.

A shorter time period than monthly will result is a more useful product. Daily will a be a better objective for future developments, although it will require a different selection of products, possibly more based on the "physically" based diagnostic ET products, where daily is a common time scale. I would suspect a more complex merging exercise, given the larger amount of time variability that needs to be captured by the merge product. ET and FluxNet datasets

Tier 3 (i.e, Greenland and Antarctica) is just a very close weighting of an obsolete version of GLEAM and a newer version. It seems a bit awkward to distribute that as part of a synthesis product. It may have been better to just remove those regions from the synthesis product given that nearly no one dares to estimate ET over there (understandably).

Some of the ET datasets considered are based on algorithms that estimate separately interception, evaporation from the canopy, and evaporation from the soil (e.g, MOD16, PT-JPL, and GLEAM). Under the assumption that routine EC observations perform very poorly for rainy conditions, in principle interception is not captured by the tower observations. In some recent ET evaluations of these products the tower data has been filtered to remove rainy periods and the interception component has not been evaluated. A discussion about this could have been interesting, given that, as far as I can see, the tower data is not filtered for precipitation conditions, and the merged product is a total ET product.

Regarding the energy closure issue, the text may give the impression that fluxes correction is always possible, but a large number of stations do not measure Rn and/or G. Given that only corrected fluxes are used in the study, I imagine that a number of sta-

tions have to be discarded as they corrected fluxes were not available (FLUXNET-2015) or could not be estimated (LaThuille-2007). This may be worth mentioning.

Merging technique

If I understand the method correctly, the weights are global (i.e. one value for all pixels) , time-invariant (i.e., an annual value), and the bias-correction is what Bishop, 2013 calls a "global bias" correction (i.e., is a single value per product using all towers in the in-sample training dataset). If this is true, if Product A was performing better than product B over some biomes and/or at some periods, the method cannot be used to weight more or less the products to reflect that difference in performance. If I am correct, I wonder if there is a way to modify the weighting to take into account those differences. We typically see that ET products perform differently at different biomes and/or seasons, so it may be advantageous to capture this in the weighting. Bishop, 2013 was quite illustrative about this. Given that for that temperature example the "truth" was quasi-global (i.e., not over a very few pixels like for the ET), just a "per-cell bias" correction, even without weighting, outperformed the weighted product with a prior "global bias" correction. Of course, the "per-cell bias" correction and weights cannot be be applied here, given the limited geographical coverage of the flux "truth", so the problem is more complex.

I guess your first take at this is your sites clustering based on vegetation type, presented in some detail in the Supplement. You conclude that it did not improve overall the DOLCE performance based on a global analysis, but I would be interested to see the results (in the Supplement I can download the figures summarizing the analysis are missing). Perhaps other schemes that better cluster the flux behavior are worth investigating for future versions of DOLCE.

Results

It is stated that the MSE plot in Fig. 3a shows the MSE of the weighted product being better than the ensemble mean, but I do not see it. The central line of the first whisker

box in Fig 3a is at the zero line. Perhaps I am missing something regarding how to read these plots. As a side note, it may be good to say what the end of the whiskers represents. In most occasions it is used to represent the max and min of the data, but it is not always the case.

In the same Fig. 3a, the whisker boxes for the individual products make me think again about the MPI product. Based on the "heavy" calibration of MPI with the same tower data used to derive the weights, I would speculate that if the MPI product was removed from the merging, the percentage improvements of this new weighted product (i.e., without MPI) over the individual products will be much smaller. This may give a different perspective of the exercise regarding the skill of the tower-based merging to combine the more "tower-independent" physically based products.

If we just concentrate on the improvements of the weighted product with respect to the equally weighted product (i.e., first whisker box in Fig 3 a-b-c), the gain in performance of the weighted product seems small. Again, if I read these plots correctly, the gain for MSE and COR is minimal, only the RSD shows some improvement. But given the definition of the RSD metric, I wonder is this is mostly associated to the bias correction. If I understand this correctly, after the bias correction mean-dataset and mean-observation will be equal, so RSD is abs(sigma-dataset –sigma-observation). In other words, I am wondering about a comparison of the equally weighted product, but with a bias-correction first, and the weighted product. I am assuming here that the equally weighted mean did not involve a prior bias correction, as nothing was stated in the paper, but I may be wrong.

In the HOM and HET comparison, I see very small improvements in MSE, larger for RSD, and not much for COR (the median of the whisker box is for MSE and COR is at the zero percentage line). And I wonder if the separation into HOM and HET sites may have also implied a separation in land covers, so the improvements we see are more related to the weighted product working better for some specific biomes. One may think that land covers such as forested areas are more likely to be represented in

the HOM class, compared with e.g. croplands. I wonder if this has been checked, i.e., that the biome representation in HOM and HET classes does not change too much.

Regarding the boxplots of Figure 5, it is true that the end of the whiskers are larger for the HET sites, but the RMSE and correlation median and percentiles look slightly better for the HET class. I think this is also worth discussing, as it can potentially indicate again that the differences in performance between HOM and HET classes are small, with just a few HET sites having bad statistics. Given that this is based on one site out-of-sample, I wonder if the bad performance at some individual sites may be nothing to do with homogeneity, but with the fact that the site is one of a kind, so any weights derived from a sample without that site are not informative. Given the location of some the sites, it would not be a surprise.

MPIBGC is the larger contributor to DOLCE, with a weight close to $\sim$0.5. Perhaps it is not surprising than the differences of MPI with DOLCE sown in Figure 6 are smaller than with LandFlux-EVAL-Diag in Figure 7, which is not part of DOLCE.

I acknowledge that it is quite difficult to come out with the right classes to try to illustrate the reliability of the weighted product. But when I look at figure 8-c what I broadly see is that arid places, Greenland and Antarctica have low reliability, snowed places medium, and the rest high. Perhaps just the uncertainty over the mean ET would have been more informative.

Discussion

The discussion about a possible selection of HOM sites to construct DOLCE is quite appropriate. Being very strict, and considering that the typical tower fetch is of the order of hundreds of meters, while the grid cell has an area of $\sim$2500 km2, possibly none of the sites will truly qualify as HOM. But if we want to keep using tower fluxes, we need to live with this.

Perhaps a simple measure to reduce the effect of this tower-fetch and cell-scale mis-

match is to try to work at finer resolutions in the future. GLEAM, MOD16 and Zhang, 2010 are already at resolutions 0.25 deg. PT-JPL will be soon available at 5 km. Perhaps a couple of products will never be available at 0.25 deg (e.g., MPI), but they could be downscaled to 0.25 deg and merged with the other products (e.g., by a nearest-neighbor interpolation if we do not want to add any extra information).

The lack of success of any clustering of the sites based on vegetation, climate, etc, possibly is more indicative of the limitations of the tower flux, rather than limitation of the conceptual idea. Even if not terribly successful, a look at how the relative weights of the different products change for different clusters can be informative regarding the products performance for different conditions.

As I mentioned before, using MPI as part of the weighted product is perfectly valid. But I think the main feature defining MPI for this study is not the fact that it is a statistical product, but more that it is a global extrapolation of the same tower fluxes used to derive the weights. As clearly stated in the paper, the tower fluxes have limitations, which will have an impact on the weighted product. Presumably, the MPI product will suffer from the same limitations than the tower fluxes, but these limitations will not become apparent when looking at differences with the tower fluxes, so it will not be reflected in the weights. An example could be the ET estimation for those biomes and moments when interception can be a relevant component of the fluxes. If the tower fluxes are not properly capturing this component, the physical methods that try to do so (e.g., GLEAM, PT-JPL, MOD16) may be penalized in the weight derivation, compared with MPI, but their fluxes may be more correct. Of course, the problem is how to merge or validate the ET products away from the tower flux data. Given that DOLCE is a global product with a relatively long time series, at least the annually integrated ET could be compared with basin-integrated differences of precipitation and runoff, which may shed more light into the merits of the individual products, the equally weighted mean product, and DOLCE.

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2017-147, 2017.