Biogeosciences Discuss.,
doi:10.5194/bg-2016-495-AC1, 2017

Biogeosciences Discussions

Interactive comment on "Water, Energy, and Carbon with Artificial Neural Networks (WECANN): A statistically-based estimate of global surface turbulent fluxes using solar-induced fluorescence" *by* Seyed Hamed Alemohammad et al.

Seyed Hamed Alemohammad et al.

sha2128@columbia.edu

Received and published: 11 February 2017

The authors proposed a new global product of GPP, ET and H by using ANN. The manuscript is well written and the topic falls on to the scope of the journal. I do have several concerns.

Re: We thank the referee for his/her positive comments.

First, the authors highlighted the use of SiF as input data. I see SiF was the

only input data related to vegetation. Therefore, with/without SiF in WECANN must give different flux estimates. What happens if the authors use EVI or NDVI instead of SiF? Any significant difference in WECANN performance?

Re: We agree with the referee's point on the evaluation of no-SIF retrieval. Therefore, in the revised manuscript we included comparisons with an artificial neural networks retrieval that has either NDVI or EVI as input instead of SIF. We comment on the differences and similarities, and why SIF is a better input for this retrieval, in particular highlighting the differences in terms of vegetation structure impact on SIF and the impact of saturation of vegetation indices (especially in forested areas and agricultural regions).

Second, what is the significant contribution from this work? Spatial (1 degree) and temporal (monthly) resolutions are too coarse. The approach is on the similar family of other machine learning methods (e.g. see Tramontana et al 2016 Biogeosciences). Stress the novelty of this manuscript. If there is any new discovery, then highlight it.

Re: There are two major new contributions in this study:

1. Using remotely-sensed SIF to estimate surface fluxes.
2. Using a machine learning algorithm (in this case artificial neural networks) to estimate fluxes from remote sensing observations at global scale.

The Tramontana et al 2016 paper uses a regression model to upscale fluxes from FLUXNET observations. However, we use remote sensing observations to estimate fluxes, and use FLUXNET towers to evaluate the performance of our retrievals. Therefore the strategy is pretty different compared to the Tramontana et al retrieval. In addition, our main objective is to show that SIF provides useful information on the rates of photosynthesis and evapotranspiration. To our knowledge this is the first direct es-

timate of fluxes based on SIF data. We revised portions of the text in the introduction section to make sure the novelty of our approach is clearly stated.

Third, the authors used MPI-BGC product as a training dataset while testing the product against FLUXNET data. As MPI-BGC product was trained against FLUXNET dataset, the approach is self-correlated. Why not evaluating the product against independent datasets from MPI-BGC? E.g. water balance derived ET in basin scale.

Re: That is not exactly correct. We train our algorithm against a target dataset which is derived from three products (including MPI-BGC) by using the Triple Collocation method and assigning *a priori* weights to every product in each pixel. This means that our target dataset has collective information from all three products and not just MPI-BGC. Therefore, we acknowledge that there is some information carried from FLUXNET tower data in MPI-BGC to our training process. However, the degree of self-correlation is mainly true for the comparison of FLUXNET-MTE with the tower estimates but not for the WECANN estimate. Indeed, it has been shown (see Jimenez et al. 2009 for instance) that the spatial and temporal correlations of a global artificial neural network are not due to the initial training dataset but to the remote sensing observations used as input.

Jimenez, C., Prigent, C., & Aires, F. (2009). Toward an estimation of global land surface heat fluxes from multisatellite observations. *Journal of Geophysical Research-Atmospheres, 114*(D6), D06305.

Moreover, given that FLUXNET-MTE uses flux tower estimates for the retrieval we would have expected this product to be the better one when compared to local eddy covariance data. In fact, we show the opposite. WECANN informed by direct remote sensing observations typically outperforms FLUXNET-MTE, especially in terms of seasonal cycle, further emphasizing the information content provided by remote sensing data.

Finally, conducting a water balance analysis will be informative while it has its own challenges because of uncertainties in other inputs for the water balance, in order to close the budget as multiple sources of information needs to be used. However, we believe that this is beyond the scope of the current study which is solely focused on developing the retrieval algorithm, also the other referee commented on the length of the paper asking us to reduce it.

Fourth, the spatial domain should be clearly defined. The authors said it is global product, but it did not include Antarctica and Greenland. Given the coarse resolution (100 km), most islands are likely uncovered but the global map (Fig 2) showed fluxes in some islands. How did it happen? Also, how to treat with water fraction for each 1-degree pixel?

Re: Thank you for the comments. We have now revised the description in the introduction section to clearly note what the coverage of the new product is.

Fifth, I recommend showing global uncertainty maps for GPP, LE, H. I think one of strengths in WECANN is its ability to quantify uncertainty. Show the uncertainty map and discuss where and why uncertainties are high. Also quantify uncertainties in global values (e.g. XXX PgC yr-1 +- Y PgC yr-1).

Re: In the revised manuscript we now include uncertainty estimates based on errors in the input data propagated into the network. We report a global average value as error is spatially and temporally variable.

Sixth, test global more carefully. When I look at Fig 2, I found higher ET in mid to south east South America (e.g. cerrado) compared to other global ET products. Also, your ET in this region is relatively very high compared to your GPP map. So, water use efficiency will be very low in this region, which is unlikely. See global distribution of C4 maps. Higher proportion in C4 in this area is likely to lead higher water use efficiency. It is notable that your ANN did not consider C4 information.

Re: The referee's point is an important one. The SIF relationship with GPP will likely change in C4 plants. However, we explicitly did not want to impose the C4/C3 (or even CAM) delimitation in the artificial neural network as it would be highly dependent on the quality of the classification map used and might be time varying. Given that we do not have partitioning of transpiration to total ET, it is hard to say whether the water use efficiency is indeed low or if rain re-evaporation and soil evaporation is the main process explaining the difference. We have nonetheless added a comment in the text emphasizing the referee's point(s).

Specific comments:

P6: why only 21 FLUXNET sites were used? More than 150 sites data are open to public

Re: We had selected this 21 sites to represent a range of climatic conditions along a geographical gradient for validation of our retrieval. Presenting evaluation metrics and temporal time series for 150 sites would lengthen the manuscript and make it hard to read. However, in the revised manuscript we will present summary statistics from comparison of WECANN retrievals against a much larger number of tower data from the FLUXNET 2015 dataset in the Appendix.

P6 L23-24: The authors explained that target data is used for training, validation, and testing. I am confused with the terminology of validation and testing. How do they differ? Also, in L36, "after training, ... was evaluated". Here, does "evaluation" indicate validation or testing? I recommend clearly defining each term, and use them consistently across the whole manuscript.

Re: We apologize for the confusion. The training, validation and testing proportions are related to the training phase of the retrieval. The back propagation algorithm uses a portion of the training data for training (basically estimating the weights of each neuron), and other portions of the training data for validation and testing that aims at checking the convergence of the training step. While after the training is done, we use a subset

of data that were not used in the training process for evaluation. We revised the text in the new version of the manuscript to clarify these terminologies.

P6 L30: NN ->ANN (?)

Re: Our apologies, this has been corrected in the revised manuscript.

P7 L9: Please define "multiple datasets." Is this training dataset?

Re: This refers to the three products that we use (together with error weights from Triple Collocation) to define a target dataset for training. We revised the text in the new version of the manuscript to clarify this.

P7 L12: What is "this" in "this prior distribution"?

Re: It refers to the pseudo Bayesian training mentioned in the lines before. We revised the text in the new version and clarified the point.

P8 L20: Is this "target estimate" from 3.2?

Re: Yes, this is the same. We made changes to section 3 and 4 of the manuscript in the new version to clarify all these terminologies.

P8 L22: Add another unit for GPP as PgC yr-1, which could be easily compared to the other studies. Same for LE (km3).

Re: Thanks for noting this. We included the new units along with previous ones in the new version of the manuscript.

P9 L29: I was surprised to see the reduction of GPP in the Saharan Desert after removing SiF. How to interpret this as we know there must be zero GPP? Also, exclusion of SiF in LE made mixed tendencies in this region. As we are confident LE and GPP are close to nil in this area, it will be interesting to test the impacts of inclusion/exclusion in SiF on LE and GPP here.

Re: This observation is true, and is caused by noise. As noted correctly by the referee,

LE and GPP are close to zero in this region; therefore, the difference between the two retrievals (with and without SIF) divided by the small amount of flux in this region is on the order of noise in the retrievals. While the percentages of change are notable, the absolute values of difference between the two retrievals are less than 2 w m^{-2} for LE and less than 0.7 gC m^{-2} day^{-1} for GPP. In addition, the noisy pattern does not show up in the H retrievals in this region. This is also another sign that the change patterns in LE and GPP are due to noise. We now emphasize this point raised by the referee in the newer version of the manuscript.

P10 L3: All three R2 looks too similar, so it is hard to tell 0.96 is higher than 0.94.

Re: Yes, we agree and have typically highlighted with bold fonts all comparable products for a fair comparison. This is further emphasized in the text of the revised version of the manuscript.

P10: The authors compared WECANN to FLUXNET-MTE, ECMWF, GLEAM and MODIS-GPP which were the training data for WECANN. I feel there should be self-correlation, so I am curious whether this is a reasonable approach.

Re: We feel that there is a misconception here. We did not compare WECANN to these products. Indeed, we are comparing WECANN and the other products against FLUXNET individual tower observations and reporting the performance of each one of them. We show that WECANN has a better performance while it is trained on the collective set of data from these products, and we believe this is an advantage of our approach.

P10 L8: I know there are few eddy flux tower data in India, so FLUXNET-MTE might involve higher uncertainty. However, this is the same situation for WECANN as it used FLUXNET-MTE and others, which are all uncertain as training dataset.

Re: It is true that there are few towers in India, but our retrieval does not rely on regional

towers to estimate surface fluxes. We train an artificial neural network algorithm using the three training products that is mentioned in the manuscript for all the pixels. That means we use the information from all the pixels over the globe to train one retrieval algorithm. This algorithm uses remote sensing observation at each point in time and space to retrieve surface fluxes. Therefore, lack of FLUXNET towers in any part of the globe would not impact the accuracy of WECANN retrievals, while this can be an issue for products that upscale tower-based observations to estimate fluxes across the globe.

P10 L4: Be quantitative. Report bias.

Re: Thanks for noting this. The point we have raised in this line (on the spread of scatter plots) can be quantitatively compared using the RMSD value that are provided in the figures. In the revised manuscript, we report this in the text as well to support the argument we are making.

P10 L20: Define "G"

Re: Corrected in the revised manuscript.

P11: Many contents in this page should move to Methods.

Re: We agree that some of the information on the description of FLUXNET data used here might be relevant to section 2 on "Data". In the revised manuscript, we re-organized the text and moved these contents to section 2.3.

P12 L5-6: Then why not removing this site given obvious deficiencies?

Re: We believe that it is informative to include this comparison, and show that the representativeness area can be a challenge in comparing large-scale remote sensing based retrievals to point based tower data. In this case, we have good knowledge of the site and its surrounding region so it is possible to investigate if the tower data is representative of the larger scale fluxes.

P12- : As the authors well recognized, I feel it is odd to compare 1 degree WECANN to several hundred meters in flux towers. All discussion from this comparison seems too subjective. I think "validation" of 0.5-degree product is unlikely possible. As your products are too coarse, I would recommend evaluating at larger scales. For example, look interannual variability of global GPP (PgC yr-1), ET (W m-2), and H (W m-2) and compare to atmospheric inversion estimates. Test whether your product could capture big climate extreme events such as Russian heatwave, Texas drought etc. Compare to other existing global land surface products which were not used as input/training dataset in WECANN.

Re: We would like to emphasize that any new retrieval algorithm development requires some validation against ground truth observations. In fact, other reviewers wanted to see some comparison. While there is some caveat in validation against point based tower data, these are the only ground based observations available for such a validation. Moreover, in the comparison against tower data many large scale variabilities such as seasonal cycle are comparable to pixel based retrievals. This is also the case for interannual variability, and we have discussed them in detail, in section 4.4 of the original manuscript. For instance, the phenology has a strong impact on the seasonal cycle of the fluxes and is here clearly highlighted when comparing the different products to flux tower estimates.

In the revised manuscript, we highlighted this limitation in section 4.4, while noting that comparison against ground-based tower observations is common practice and is what the community indeed looks for when a new retrieval algorithm is developed. We believe that specific drought or flood events would lack the generality provided here when comparing all years/months.

Interactive comment on Biogeosciences Discuss., doi:10.5194/bg-2016-495, 2016.