Biogeosciences Discuss.,
doi:10.5194/bg-2016-495-AC2, 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

This manuscript is well written and deserves consideration for publication in this journal. However, I have the following issues that need to be addressed.

Re: We appreciate the referee's positive feedback and provide responses to his/her comments below.

The paper proposes an empirical machine learning 'meta-model' to try to learn

from different existing datasets to combine their strengths and factor out their limitations. On one hand, I appreciate this effort to bring together different datastreams and somehow harmonize them through this new consolidated product, but on the other, I am wary of this approach of blindly adding further algorithmic layers without really trying to understand mechanistically why the initial datasets have shortcomings. If all products are equally off in some parts, combining them just gives the false impression we are going in the right direction while reality is still off. Also, the FLUXNET-MTE used as training is already a machine learning product driven by various input variables, very much like WECANN is. Furthermore, there is quite some circularity in the work since the FLUXNET-MTE and MODIS GPP are both strongly based on the same fluxtowers used here for validation. I deem that all these points need to be acknowledged clearly and discussed thoroughly.

Re: We acknowledge this concern, and would like to bring the following points to the referee's attention:

1. The WECANN machine learning retrieval is quite different than FLUXNET-MTE in the sense that we use remote sensing observations and estimate surface turbulent fluxes while FLUXNET-MTE upscales tower-based observations to estimate surface fluxes at global scale. Although both approaches use machine learning techniques (artificial neural networks in the case of WECANN and regression in the case of FLUXNET-MTE) their retrieval algorithms are quite different and directly informed by only remote sensing observations in WECANN, which we believe is an important means of better constraining the retrievals.

2. Moreover, our training approach uses all the spatial and temporal observations during the training period (2008-2010) to develop one single neural network for the global retrievals. This network is then used with remote sensing observations as input to retrieve surface fluxes. Therefore, if in a few percentage of times and

pixels, all the three training products are equally off this will be mitigated by the larger number of pixel/time data points that have more accurate estimates in other places and other times. In addition, the network can even correct the seasonal cycle when learning from an incorrect seasonal cycle training data, as the remote sensing inputs provide the information on the seasonal cycle directly. This has already been demonstrated by previous studies such as Jimenez et al. (2009).

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.

3. On the issue of validation against FLUXNET tower data, we acknowledge that two of the training products use FLUXNET data for their calibration or as input but virtually all products have been calibrated in some ways or tested against eddy-flux tower. It does not however mean that the products are not independent: indeed the products are typically calibrated to reproduce either the annual mean or are adjusted per season at very few sites but not the exact temporal structures of the eddy-covariance observations (except for FLUXNET-MTE). Here, we use the three training products together with *a priori* weights calculated from Triple Collocation to define a target dataset that has collective information from all three of them. Then, we train our network on the target dataset. Finally, we validate the retrievals of WECANN against FLUXNET tower data and compare its performance with the performance of the three training products. While some information from FLUXNET observations propagates through the training products to WECANN training, the comparison results against FLUXNET observations show that WECANN learns from the three products collectively and performs better than any of them individually, emphasizing that our strategy works well. In addition, it is clear that WECANN does not have the seasonal biases seen in most retrievals (see e.g. FI-Hyy site where WECANN correctly captures this cold region's photosynthesis and evapotranspiration compared to the other products).

We also made changes to sections 3 and 4 of the manuscript in the revised version to better reflect on these points.

Could you specify why you use the SYN products (Level 3) from CERES instead of the EBAF ones (Level 3B)? The later have been energy balanced according to the product specifications. Wouldn't this be an advantage in your case?

Re: Our goal here, as also mentioned in section 1 of the manuscript, is to only use remotely-sensed observations as input. The EBAF product is based on a model with some remote sensing observations; therefore, we decided to use the SYN product to avoid any model addition.

In the construction of the ANN, I would welcome to have some justification of why tangent sigmoid transfer functions are used instead of linear ones. I know this is often done, but it seems very arbitrary.

Re: In this case, we tried the tangent sigmoid (the common choice) as well as linear, and did not see any notable changes in the performance of the network. Therefore, we chose to use the typical tangent sigmoid function. This has been added to the text .

Also, I did not quite understand how the 20% of 'testing' data is used. I clearly see that 60% are used for training and 20% for validation, but how exactly do you use the other 20%? Perhaps this just needs some rephrasing in the text for clarification.

Re: We apologize for the confusion on this matter. In the revised manuscript, we explained this in more detail. In summary, these percentages are for the data that are used in the training process. This is standard practice in artificial neural networks training. The back-propagation algorithm uses the training portion of the data for estimating the weights of neurons in the network, and the validation and test data are used to evaluate convergence of the training. These are separate than the data that we used for validation later on. Our validation, uses a subset of data that are not used in the

training, to make sure the network is not over-fitted to the training data.

Comparison with fluxtower measurements is not appropriate as the difference in spatial support is just too different (1 squared degree vs <1km2). Saying that WECANN performs better that other products based on individual towers while all these products cover such a larger area (by several orders of magnitude) just does not make much sense (even if it has been done in other studies). The authors would need to do some filtering of the towers to select only those that can be considered representative (e.g. http://doi.org/10.1016/j.rse.2016.04.027), although I doubt this would leave many valid towers for pixels of 1 squared degree. Another option that may be more feasible would be to make an evaluation at a larger aggregation scale, such as for clusters of similar climates and plant functional types. Making such averages from the fluxtowers on one side and from all pixels that are comparable in this respect on the other would reduce the number of measurements for validation, but would render them more credible. I would also suggest to exploit more of the available towers in the Fluxnet2015 dataset instead of only 21.

Re: We acknowledge that comparison against point based tower data has its own limitation (as we also have noted in the manuscript), but these are the only ground-based validation data that is available for evaluating a new global product. For this reason, we used a selection of sites spanning a geographical gradient and provided detail explanation on the comparison results in each site based on the knowledge of the land cover / land use around the site to make sure the differences, if any, can be explained either by uncertainties in WECANN retrieval or representativeness of the of the towers. As the referee notes, filtering the towers based on representativeness might not leave us with any tower to use.

In the original manuscript, we only used 21 towers that were selected to represent a wide range of climatic conditions and we would be able to explain the results of each one of them, including the time series plots in detail. However, due to the request of

both referees we will include summary statistics from comparison of WECANN against all the sites in FLUXNET 2015. We also want to empathize that some features of the flux tower observations such as phenology and seasonality are correctly picked up by our retrieval compared to other products and are only moderately affected by the heterogeneity within the pixels (except if there would be a very different even composition of deciduous and conifers for instance). We also want to emphasize that this is the first retrieval using SIF and directly assessing its usefulness for flux retrievals. For those reasons, we believe that using coarse resolution algorithms does not alter the interest of the study.

The part pretending to demonstrate the value of SIF is also inadequate as the authors only test the effect of removing this one input. By doing so, any information of the actual vegetation phenology is lost, which would necessarily reduce the performance. What would be interesting would be to show that SIF provides better information that the classical vegetation indices like NDVI or EVI. To do so, the SIF input of the ANN should be replaced by one of these and then a judgement on the pertinence of SIF can be made.

Re: We appreciate referee's comment on this point. In order to better demonstrate the value of SIF observations we have included comparisons with retrievals that have only NDVI or EVI instead of SIF in the revised manuscript. This better shows the value of having SIF as an input in retrieving surface fluxes. Thank you for this important comment. The results further emphasize the difference between SIF and purely vegetation structure and phenology (as well as saturation effects of vegetation indices).

Finally, the manuscript is often too long and too descriptive in several parts describing the graphs and maps. This needs to be reduced drastically. Most of what is being said can be easily inferred from the reader by looking at the graphs, while deeper discussion on why discrepancies occur between products and fluxtowers would be more welcome. Also, please remove the extensive references to different parts of the text and the description of the structure of the

paper (e.g. page 3 lines 10-20), I think they are lengthening the text needlessly.

Re: Given the novelty of the approach we feel that it is important to correctly describe the different steps of the analysis as many are relatively new such as the machine learning and the triple collocation. We had received the opposite comments before that we were not sufficiently describing the details; hence, why the article goes into the details of the retrieval.

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