Dear Editor,

Please find below a point-by-point response to the reviews and a marked-up manuscript version. We have substantially changed the original manuscript to accommodate the demands from the second reviewer. Major changes are: (1) better explaining why we are limited to this sample of three ET products; (2) working with product anomalies to address the issues related to biases between the products; (3) implementing an inverse error-variance scheme to weight the anomalies, including the use of the full error covariance matrix to address the data sets dependencies; (4) better describing the NN methodology; and (5) more clearly stating whether for this particular exercise the weighted average can be considered advantageous compared with the simple average. Moreover, the readability has been improved by accommodating minor textual edits throughout the document. To guide the reviewers through the manuscript we have marked up specific request in red, clearly indicating the page and concerned lines in the response to the reviews, and adding short sentences in red in the manuscript to signal major changes to the text.

Sincerely yours,

Carlos Jimenez (on behalf of all co-authors)

Reviewer 1

The authors have successfully addressed my comments on the previous version of the manuscript. The extension of the number of merged models as well as the extrapolation to the global scale improved the value of the analysis.As such, I only have minor comments/corrections at this stage. Note that the line numbers in the following refer to the file hess-2017-573-manuscript-version5.pdf.

We thank the reviewer by going again through the manuscript and providing us with comments to improve it.

Specific comments:

line 205: how was the cross validation performed (leave-one-out)?

R: The cross-validation mentioned here is a standard technique to prevent the NN to over-fit to the training dataset, and it is applied here any time the NN is trained, no matter which training dataset is used. The leave-one-out is a different strategy, having the objective of assessing the quality of the dataset in terms of capturing the general distribution of the data. In

practical terms we test this by removing one station from the training data set, and checking if the weights of the station left aside can be properly replicated by the NN. If it is not the case, this is a strong indication pointing towards training dataset not robust enough to represent the all-conditions true distribution. These two options are now better discussed in P7-L209-227 of the revised manuscript, with the results of the tests presented in Section 7.3.

line 212: remove one of the double "the"

R: Corrected.

line 218: wrong placement of brackets for Willmott (1982).

R: Corrected.

line 240: semicolon after "Release 3.1"

R: Corrected

line 321: "World" instead of "Wordl"

R: Corrected

line 392: delete "also" (the previous station showed large inter-annual variability, not seasonal variability)

R: Corrected

Figure 7 caption: "axes" instead of "axis" in the last sentence

R: Corrected

Figure 8: adjust the size of the lower right panel to match the other panels.

R: Done.

line 514: "first" instead of "firs"

R: Corrected

line 574: missing "and" in "0.93 (MSWEP) 0.88 (WorldClim)"

R: Corrected

line 577: "catchment water budget" sounds more commonly used to me (here and some other instances)?

R: Replaced.

line 631: "In" instead of "Tn"

R: Corrected.

line 634: "factor" instead of "factors"

R: Corrected.

Reviewer 2

The authors have expanded on their original submission, by extrapolating the estimated weights using a neural network, and adding a third product to the merger. While the manuscript is certainly improved from the first submission, I still have major reservations about the motivation for this work, and the appropriateness/accuracy of the chosen methodology. These issues would need to be address prior to publication.

We thank the reviewer by going again through the manuscript and providing us with valuable comments to improve it. We have better motivated our work, updated the methodology to work with anomalies, and more clearly stated the conclusions of our exercise concerning the adequacy of using a weighted or unweigthed average for this particular combination of products.

1.
The motivation for merging these particular products needs to be provided in the manuscript (not just in the reviewer response). Also, there are other products available that you could use, say from reanalyses.

R: We are rephrasing this part of the introduction to more clearly state that this merging effort is within a specific ET modelling framework: the WACMOS-ET project. We are also mentioning that as far as we know there are no other daily global ET products publicly available. All this is added in P3-L58-68.

Concerning the reanalysis, their surface fluxes are a different type of product than the ones targeted here. They are normally classified as "derived" quantities, as their estimates are not directly constrained by the observations, and we do not intend to merge them here with the more observation-driven fluxes of the WACMOS-ET project. We have added an statement in the introduction to highlight this (P2-L43-48).

Given that all three input ET products are forced with the same data, it is not clear to me why using a statistical method (based on random errors) would be beneficial to account for systematic differences (driven by model structure / parameters) between the three products. This is a fundamental flaw in this work.

R: The three ET products share some common forcing data (namely the surface radiation and the air humidity), but they also use different products (e.g. the vegetation products and precipitation only being used by GLEAM). This is described in Section 3.1. Therefore, the differences are not only driven by model structure/parameters and the difference between the gridded products and the tower ET contains both random and systematic components, as

shown in the original Figure 11. Below, we discuss further why our merging methodology is adequate after updating it as suggested by the reviewer (see below).

While this is not actually stated, it seems to me that the main point of this work is to determine whether these three products can be improved upon by merging them, and if so, can they be merged more effectively than using an (unweighted) average.

R: We already stated in the introduction: "As such we pose the question: can a product combining the GLEAM, PT-JPL, and PM-MOD estimates result in a more accurate ET estimate? To make the paper objective clearer we are now also mentioning the simple average in the introduction, and that we are comparing it with the weighted product (P3-L76-79).

There are two steps to this:

i. Are the weights (or more importantly, the final product) significantly different between the standard averaged, and with the eights calculated using a more sophisticated approach?
Currently this has not been adequately answered, due to the flawed method used to estimate the weights.

R: See our comments below regarding our changes to the methodology.

ii. Can the weights calculated at the limited number of tower locations be usefully extrapolated globally?
Currently this cannot be answered, as not enough information is given regarding the NN used to extrapolate the weights.

R: See our comments below regarding the application of the NN.

2. As in the first review. The inverse-error variance weights are based on the assumption of unbiased and independent data sets. The data sets used here are strongly dependent and biased, and this cannot be ignored. At a minimum for publication, the weights (and consequently the averaging) needs to be based on anomalies (to remove the bias, and the more systematic aspects of the dependence), and the lack of independence needs to be acknowledged prominently, including qualifying all conclusions by noting that these data sets were not independent. If you don't have enough data to use anomalies, then you cannot apply this method by ignoring the inconvenient biases.

R: We fully agree with the reviewer that an inverse-error variance merging requires unbiased estimates. Only under those conditions this statistical method can be considered and optimal estimator, with the resulting merged product minimizing the error variance. Therefore, strictly speaking, if we consider the *in situ* observations unbiased with respect to the truth, the products should be debiased against the *in situ* observations. This kind of debiasing will strongly constrain the merged product to the tower absolute values, and would require a global estimation of the bias for the final product. In essence, this would be trying to reproduce the fluxes, as in Yao et al., 2017. Note that this is not the objective here, but the intention is to weight the original gridded products giving more weight to the products that reproduce the tower variations more closely.

After some tests we finally implemented a new merging scheme that allows working with anomalies, but without correcting to the tower absolute values. In short, we first derive time series of anomalies for each product and the tower observations by removing their individual seasonal climatologies. Then, we calculate the full covariance matrix of the difference of the gridded product anomalies and the tower anomalies as the basis for the weighting, so we account for the expected product correlations due to modelling/forcing similarities. Finally, we weight the anomalies and add them to the climatology of the simple-averaged product. This is now fully explained in P5-L155 to P6-L186.

3. The NN network is not adequately explained, and reads like it has been blindly applied rather than carefully investigated. As with the first review, I still have concerns about over-fitting. Since so little information is given, it is not clear whether the NN was not useful for extrapolation, because of the limited locations of the tower observations, or because the NN itself was inadequate.

R: We have ample experience applying a similar NN setup for different geophysical problems (e.g., Jiménez et al., 2003, Jiménez et al., 2009, Jiménez et al., 2012) and are certain that the NN is adequate for this application. The NN description was short as we firmly believe that the limitations here are coming from the data set, not from the method used to find the statistical relationship between the weights and the model inputs. Multi-layer perceptrons can be trained with different backpropagation algorithms, and their generalisation capacity can be optimised by techniques such as structural stabilisation, cross-validation, or regularisation. In our experience all these different NNs perform very similarly when properly set up. This has been further explained in the text (see below).

For publication, the NN needs to be more adequately explained. This includes details of how the training data sets were split into training and evaluation data within the NN algorithm (and how this then relates to the experiments where a single station was removed), how the input data was selected / what else was tested, and how you ensure that the output weights equal 1.

R: Full details are given now at P7-L197-227. It is worth indicating that controlling the optimal complexity of the model underlined by the NN is relatively easy by using some of the techniques mentioned before, so we are confident that we are not over-fitting in that sense. However, this does not tell much about the "quality" of the data set to truly sample all possible conditions, which will be always a concern when using the existing pool of tower data for global applications. Therefore, the typical stratification of the training data sets to test whether the available sampling of the true distribution is adequate to capture most possible conditions.

Also, it isn't stated whether a separate NN is calculated for each day of the year, or if all data is thrown in to a single NN. I suspect it is the former, but would strongly recommend the latter.

R: Yes, the NN is trained to model the annual statistical relationship between the weights and model inputs. We are making that clear now (P6-L191-193). We would not advice to model the daily relationship due to the resulting very small database for the individual daily trainings, and the suppression of the temporal variability in the daily training data sets. Both

things do not help define the link between variations in the NN inputs and weights, and are likely to impact the robustness of the found relationships.

How do you deal with hemispheric differences? It doesn't really make sense to predict SH weights for a day of year, based on NH data only (there is only one SH tower, in Australia).

R: We agree that this is a severe limitation for a global extrapolation. Still, we think it is worth presenting a global extrapolation, but testing if the biomes covered by the stations in the NH could be somehow representative of the conditions in the SH. This is a gross assumption, and our tests indeed showed that it is not the case. A test about this can be found in P25-L543 to P26-L560.

By training a single NN using data for all days (and not including day of year in the training data) you could withhold data for some days to test whether the NN can at least reproduce the sites that are sampled.

R: A similar test is now included in Section 7.3 and described in P25-L517-542. The "all stations" statistics are derived from days of the year not seen by the NN during the training phase. For that, from the original stations data set 15% of the available days at each station are removed before training and used for the statistics.

MINOR:

1A general grammar check would be useful, with attention paid to missing possessive apostrophes.

R: The whole manuscript has been carefully edited paying attention to the grammar.

L49: please double check that the MTE product is a regression. I thought it was a machine learning approach, but could be wrong.

R: Regression predictive modelling, i.e. is the task of approximating a mapping function (f) from input variables (X) to a continuous output variable (y), is a standard technique in machine learning approaches. Machine learning approaches are typically applied to either regression or classification problems.

Section 2.1 (from first review also). Add the spatial resolutions in here.

R: Added.

Replace all instances of 'inverse-variance weighting' with 'inverse error variance weighting'

R: Corrected.

L210. "Only the systematic component of the error can potentially be captured by the NN, and there is not warranty that all the systematic errors are dependent on the model inputs". This is incorrect/misleading. "systematic errors" usually refers to biases, and

these methods can predict the mean square of the random errors (which is what your inverse error variance weights should reflect). Note that statistical methods like this do not predict individual errors, rather they predict the tendency towards larger/smaller errors.

R: Yes, the sentence is indeed misleading, and we have rephrased it. Apart from that, we assume that by statistical methods the reviewer refers to the NN. If that is the case, we are certainly not trying to predict individual errors, but to model the distribution of the weights, conditioned to the ET model inputs and model ET estimates. For this, the NN is calibrated by minimizing the sum of square errors, where each error is the difference between each target weight of the training data set, and the NN prediction.

Regarding the second half of the sentence, this is why you need to do some work to show that you have selected appropriate input data sets for NN.

R: We agree that this part of the sentence was again not clear. We meant that the NN is used to model the statistical distribution of the weights, with this distribution not only depending on the variables used as predictors in the NN approach, which means that we can never perfectly predict the weights. This has been added to the manuscript (P7-L194-197).

L273: quantify 'too close'. How is "clearly not representing the overall land cover" determined?

R: We agree with the reviewer that this is rather subjective. In this context, "too close" meant that a water body could be visually identified in an aerial picture of the 25 km cell surrounding the eddy-covariance tower. The same visual inspection was used to discard stations where the station surroundings were clearly not representative of the station cover. This has been added to the manuscript (P10-L285).

Explicitly note here that the towers do not provide comprehensive global coverage, and don't cover many biomes, climate regimes, or the Southern Hemisphere.

R: We believe that this was already stated in the paper, but we are further adding that there are only 2 stations in the Souther Hemisphere (P11-L290)

L280: representativity errors should be mentioned here (tower to 25 km).

R: Certainly, representativeness errors cannot be ignored given the large mismatch between tower fetch and the spatial support of satellite imagery, but this section discusses the tower errors independent of the application. We chose to place this comment at the start of the next section, when we introduce some ancillary data to investigate the tower surroundings homogeneity (P13-L331-334).

Figure 2 : 100 Wm2 is a huge range for each colorbar. Please plot with a finer discretization.

R: 100 W/m2 corresponds here to a discretization of the full ET range in 10 intervals and the corresponding 10 colours. In our view this is an adequate level of discretization for this type

of global plots, and we have already used it in other flux related publications without any issues (e.g., Jiménez et al., 2009, Jiménez et al., 2010). Therefore, we prefer to leave the plot as such.

Figure 4: It is stated repeatedly that the weights don't differ much from 1/3, and yet this plot shows a large deviation. Statistics are needed here to quantify the divergence from 1/3.

R: The weights differ now from the 1/3 value after the new merging scheme. Nevertheless, we are replacing Figure 4 with a box plot to display the weight statistics in a more informative way.

Figure 7: It doesn't really make sense to plot global fields for three month blocks.

R: The authors do not fully agree with this comment, as these maps show the average seasonal patterns of evaporation, which helps to assess whether the temporal dynamics in the datasets are well represented at the seasonal scale.

Figrue 10: This plot is difficult to understand. Why not plot Ih v. RMSE?

R: We thank the reviewer for this suggestion. A scatter plot of Ih versus RMSD is certainly also a possibility, but here we intended to show whether a decreasing homogeneity results in an increasing RMSD. Plotting like this we can also display the linear square fit of the normalized RMSD of the sorted stations, so we prefer this type of figure. We are rewriting the figure caption to make this clear.

Ih has not been defined anywhere in the manuscript.

R: We would like to point the reviewer to Eq. 7 of the original manuscript, where Ih was already defined.

L507: what about systematic errors in the obs? (inc. representativity)?

R: This is indeed an issue, and it can be problematic when the observations are used to improve model estimates. Note that possible systematic errors in the tower observations were already discussed in Section 3.2 of the original manuscript though , now at (P11-L292-299).

L509: "if the difference with the observations were mostly random in nature, we should not expect the observations to provide much guidance to combined the products". This is not correct. See comment on L210.

R: We agree with the reviewer. Only in case of using the observations to improve a particular model estimate, the systematic differences are useful to detect and correct model issues. In the framework of an inverse error variance scheme to combine estimates the variance of the random errors guides the weighting.

L521: No. You could have defined the bias over the data time period that you do have.

R: We agree with the reviewer that the bias is an issue, and we believe we properly address this now by working on anomaly time series.

L565: the occurrence of negative weights is because you have dependent products.

R: This is indeed the underlying reason, and this is clearly acknowledged now throughout the manuscript (e.g. P17-L400).

L581: Qualify that this approach assumes that the surface water storage has not changed over this time period, and that this assumption will not necessarily hold over the limitted time period you are using.

R: This has now been acknowledged at P27-L577-578.

Note that this is the same precip used in the ET products, and that this does not represent an independent evaluation.

R: We agree, and we would like to point out that we were already tackling this issue by adding a second precipitation product. This is now highlighted again in Section 7.4 (P27-L570-575).

L605: either explain why the slope is an important statistic here, or delete it.

R: Given that the expectation here is the ET from the models equalling the basin ET inferred, closeness of the slope of the linear fit to one is a desirable target. Nevertheless, with the new merging scheme the slopes of the merged product are more similar, so we remove the slope discussion and add the bias as a third statistic inference.

References

Jiménez, C., P. Eriksson and D. Murtagh, **First inversions of observed sub-millimetre limb sounding radiances by neural networks** , J. Geophys. Res.,108(24), 4791, doi:10.1029/2003JD003826, 2003.

Jiménez, C., C. Prigent, and F. Aires, **Toward an estimation of global land surface heat fluxes from multisatellite observations**, J. Geophys. Res., 114, D06305, doi:10.1029/ 2008JD011392, 2009.

Jiménez, C., C. Prigent, B. Mueller, S. I. Seneviratne, M. F. McCabe, E. F. Wood, W. B. Rossow, G. Balsamo, A. K. Betts, P. A. Dirmeyer, J. B. Fisher, M. Jung, M. Kanamitsu, R. H. Reichle, M. Reichstein, M. Rodell, J. Sheffield, K. Tu, K. Wang, **Global inter-comparison of 12 land surface flux estimates**, J. Geophys. Res., 116, D02102, doi:10.1029/2010JD014545, 2011.

Jiménez, C., D. Clark, J. Kolassa, F. Aires, and C. Prigent, **A joint analysis of modeled soil moisture fields and satellite observations**, J. Geophys. Res., DOI: 10.1002/jgrd.50430, 2012.

Yao, Y., Liang, S., Li, X., Chen, J., Liu, S., Jia, K., Zhang, X., Xiao, Z., Fisher, J. B., Mu, Q., Pan, M., Liu, M., Cheng, J., Jiang, B., Xie, X., Gr nwald, T., Bernhofer, C., and ROUPSARD, O.: **Improving global terrestrial evapotranspiration estimation using support vector machine by integrating three process-based algorithms,** Agricultural and Forest Meteorology, 242, 55–74, 2017.

Manuscript prepared for Hydrol. Earth Syst. Sci.
with version 2015/04/24 7.83 Copernicus papers of the LaTeX class copernicus.cls.
Date: 15 June 2018

Exploring the merging of the global land evaporation WACMOS-ET products based on local tower measurements

Carlos Jiménez[1,2], Brecht Martens[3], Diego M. Miralles[3], Joshua B. Fisher[4], Hylke E. Beck[5], and Diego Fernández-Prieto[6]

[1]Estellus, Paris, France
[2]LERMA, Paris Observatory, Paris, France
[3]Laboratory of Hydrology and Water Management, Ghent University, Ghent, Belgium
[4]Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California, USA
[5]Department of Civil and Environmental Engineering, Princeton University, Princeton, New Jersey, USA
[6]ESRIN, European Space Agency, Frascati, Italy

Correspondence to: Carlos Jiménez(carlos.jimenez@estellus.fr)

Abstract. An inverse error variance weighting of the anomalies of three terrestrial evaporation (ET) products from the WACMOS-ET project based on FLUXNET sites is presented. The three ET models were run daily and at a resolution of 25 km for 2002–2007, and based on common input data when possible. The local weights, derived based on the variance of the difference between the tower ET anomalies and the modelled ET anomalies, were made dynamic by estimating them using a 61-day running window centered on each day. These were then extrapolated from the tower locations to the global landscape by regressing them on the main model inputs and derived ET using a neural network. Over the stations, the weighted scheme usefully decreased the random error component, and the weighted ET correlated better with the tower data than a simple average. The global extrapolation produced weights displaying strong seasonal and geographical patterns, which translated into spatiotemporal differences between the ET weighted and simple average ET products. However, the uncertainty of the weights after the extrapolation remained large. Out-sample prediction tests showed that the tower data set, mostly located at temperate regions, had limitations with respect to the representation of different biome and climate conditions. Therefore, even if the local weighting was successful, the extrapolation to a global scale resulted problematic, showing a limited added value over the simple average. Overall, this study suggests that merging tower observations and ET products at the time and spatial scales of this study is complicated by the tower spatial representativeness, the products coarse spatial resolution, the nature of the error in both towers and gridded data sets, and how all these factors impact the weights extrapolation from the tower locations to the global landscape.

1 Introduction

The surface latent heat flux governs the interactions between the Earth and its atmosphere (Betts, 2009), is an essential component of the water and energy cycles (Sorooshian et al., 2005), and thus plays a key role in the climate system and on the linking of biochemical cycles (Wang and Dickinson, 2012). Terrestrial evaporation (ET) – the associated flux of water from land into the atmosphere – is also an important variable in the management of agricultural systems, forests, and hydrological resources. Hence, estimates of ET at different spatial scales, ranging from individual plants for managing irrigation, to basin scales to evaluate water availability, are required by many applications(e.g. Dunn and Mackay, 1995; Le Maitre and Versfeld, 1997; Gowda et al., 2008; Fisher et al., 2017).

Point-based measurements of land heat fluxes are typically conducted during field experiments (Pauwels et al., 2008) or by more permanent monitoring systems, such as lysimeters (Hirschi et al., 2017) and flux tower networks (Baldocchi et al., 2001). However, these are ultimately point measurements that require specific equipment and cannot be applied for routine monitoring over large areas. Therefore, more readily available meteorological observations are often combined with well known flux formulations (e.g., Monteith, 1965; Priestley and Taylor, 1972) to obtain regional-scale estimates.

To derive global estimates, a central challenge remains: ET does not have a direct signature that can be remotely detected. As an alternative, satellite remote sensing observations related to surface temperature, soil moisture, or vegetation can again be combined with traditional flux formulations (e.g., Monteith, 1965; Priestley and Taylor, 1972) to derive global estimates at different time and spatial scales. This has led to the raise and proliferation of satellite observation-based retrieval models (and subsequent data sets) of ET over the last few years (for overview see Wang and Dickinson, 2012; Zhang et al., 2016). Global flux estimates are also available from atmospheric reanalyses (e.g. Dee et al., 2011), but are often treated separately as they are not as directly constrained by observations as the satellite data-driven data sets (Jimenez et al., 2011; Mueller et al., 2013). In addition, the latter are specifically designed to estimate ET, and while also uncertain, their errors are in principle more traceable due to their lower complexity. Nonetheless, satellite-based ET products also show large discrepancies which are put in evidence when inter-compared and evaluated against in situ flux networks (Jimenez et al., 2011; Mueller et al., 2011; McCabe et al., 2016).

Far from discouraging the use of these ET datasets, the inter-product differences have been perceived as an opportunity to foster research and find new means to combine these datasets in an optimal manner. So far, these efforts have ranged from simply averaging a number of ET products (Mueller et al., 2013) to more complex approaches, such as weighted averages (Hobeichi et al., 2018), fusion algorithms where the original ET products are combined to reproduce flux observations (Yao et al., 2017), or integration methodologies that seek consistency between ET products and related products of the water cycle (Aires, 2014; Munier and Pan, 2014). ET products based on a direct regression of tower ET on a set of explanatory variables also exist(Jung et al., 2011).

Aiming at improving the predictive capability for ET, the WAter Cycle Multi-mission Observation Strategy – ET project (WACMOS-ET, *http://wacmoset.estellus.eu*) compiled a forcing data set covering the period 2005–2007, and ran four established ET models using common forcing to explore the uncertainties and accuracy of the underlying algorithms (Michel et al., 2016; Miralles et al., 2016). Three of the models – the Priestley-Taylor Jet Propulsion Laboratory model (PT-JPL, Fisher et al., 2008), the Global Land Evaporation Amsterdam Model (GLEAM, Miralles et al., 2016), and the Penman–Monteith algorithm from the MODerate resolution Imaging Spectroradiometer (MODIS) evaporation product (PM-MOD, Mu et al., 2011) – were run to produce 3-hourly and daily estimates at 0.25^o spatial resolution. As far as we know, they remain the only publicly available global ET estimates at these spatiotemporal resolutions.

Analyses of the WACMOS-ET estimates showed substantial differences between the three model products, both at the point scale (Michel et al., 2016) as well as globally (Miralles et al., 2016). As such we here pose the question: can a combination of these estimates result in accurate ET? The simplest approach is to assume that all products are equally uncertain, merging them with a simple average. A more elaborated approach is to assign weights to each product based on an accurate description of the specific product uncertainties. However, even if some attempts to derive model uncertainty exist (Miralles et al., 2011a; Badgley et al., 2015; Loew et al., 2016), the complexity to derive estimates of ET from remote sensing data means that reliable quality assessment is only attained through validation against tower flux measurements. Therefore, here we explore a local flux tower-based weighting of GLEAM, PT-JPL, and PM-MOD and compare it with the more typical simple average, followed by an appraisal of the potential to globally extrapolate the resulting merging framework.

2 Methods

2.1 ET models

The GLEAM, PT-JPL, and PM-MOD models, and the inputs required to run them globally at a 0.25^o spatial resolution are extensively described by Michel et al. (2016) and Miralles et al. (2016). Only the main differences with respect to the original WACMOS-ET runs are fully detailed here. Note that the original 2005-2007 period is extended here to cover 2002-2007, and that the models are only run at daily time resolutions.

2.1.1 GLEAM

GLEAM is a simple land surface model fully dedicated to deriving evaporation. It distinguishes between direct soil evaporation, transpiration from short and tall vegetation, snow sublimation, open-water evaporation, and interception loss from tall vegetation. Interception loss is independently calculated based on the Gash (1979) analytical model forced by observations of precipitation. The re-

maining components of evaporation are based upon the formulation by Priestley and Taylor (1972) for potential evaporation, constrained by multiplicative stress factors. For transpiration and soil evaporation, the stress factor is calculated based on the content of water in vegetation (microwave vegetation optical depth) and the root zone (multilayer soil model driven by observations of precipitation and updated through assimilation of microwave surface soil moisture). For regions covered by ice and snow, sublimation is calculated using a Priestley and Taylor equation with specific parameters for ice and supercooled waters. For the fraction of open water at each grid cell, the model assumes potential evaporation.

The recent GLEAM v3 model of Martens et al. (2016) is adopted here and replaces the model of Miralles et al. (2011) previously applied for the WACMOS-ET runs. Major differences related to the previous model are a revised formulation of the evaporative stress, an optimized drainage algorithm, and a new soil moisture data assimilation system.

2.1.2 PT-JPL

The PT-JPL model by Fisher et al. (2008) is a relatively simple algorithm to derive ET. It uses the Priestley and Taylor (1972) approach to estimate potential evaporation, and then applies a series of stress factors to reduce from potential to actual evaporation. The land evaporation is partitioned first into soil evaporation, transpiration, and interception loss by distributing the net radiation to the soil and vegetation components. Unlike GLEAM, the stress factors in PT-JPL are based on atmospheric moisture (vapour pressure deficit and relative humidity) and vegetation indices (normalized difference vegetation index, and soil adjusted vegetation index) to constrain the atmospheric demand for water. The partitioning between transpiration and interception loss is done using a threshold based on relative humidity, and is therefore conceptually quite different from the precipitation based calculation in GLEAM. There is no independent estimation of snow sublimation, and the same algorithms are applied for snow-covered areas.

For this study, optimized vegetation products are used as inputs to the model. In WACMOS-ET, the Leaf Area Index (LAI) and Fraction of Absorbed Photosynthetic Active Radiation (FAPAR) products, derived from the Joint Research Centre Two-Stream Inversion (JRC-TIP) package (Pinty et al., 2007, 2011a, b), were converted by a simple biome-dependent calibration to a LAI/FAPAR product consistent with the Moderate Resolution Imaging Spectroradiometer (MODIS) LAI/FAPAR before being used as inputs to the model (Michel et al., 2016). Under the assumptions that the JRC-TIP FAPAR is related to the radiation absorption by the green fraction of the canopy, while the MODIS FAPAR is more related to green and non-green leaf area, a new use of the WACMOS-ET vegetation products is proposed. First, the WACMOS-ET JRC-TIP FAPAR is assumed to be close to an Enhanced Vegetation Index (EVI), and it is scaled by the factor 1.2 to become closer to the FAPAR expected by the model, as in the original PT-JPL equations (Fisher et al., 2008). Second, the WACMOS-ET MODIS-like FAPAR is used as the Fraction of Intercepted Photosynthetic Active

Radiation (FIPAR) expected by the model, which in turn is used by the model as a proxy for the fractional total vegetation cover. Using the original relationships in the model, the fractional total vegetation cover is related to a total (green and non-green) LAI, which is then used to partition the net radiation into their soil and canopy components.

2.1.3 PM-MOD

The PM-MOD is based on the Monteith (1965) adaptation of Penman (1948), and the version applied here follows the implementation of Mu et al. (2011). It estimates ET as the sum of interception loss, transpiration, and soil evaporation. Aerodynamic and surface resistances for each component of evaporation are based on extending biome-specific conductance parameters to the canopy scale using vegetation phenology and meteorological data. The surface resistance schemes uses LAI, with further constrains based on air temperature and vapour pressure deficit, avoiding the need of soil moisture and wind speed to parameterize the resistances. Different from GLEAM and PT-JPL, which do not use tower-based calibration, some of the resistance parameters require a biome-based calibration derived from a selection of tower measurements. As for PT-JPL, there is no specific parameterization for snow-covered areas.

The WACMOS-ET LAI/FAPAR products are used with PM-MOD as in Michel et al. (2016), i.e., the model is run with the vegetation products rescaled by a biome-dependent calibration to make them consistent with the expected MODIS values. As the biome-based calibration of PM-MOD was derived with MODIS products, any errors introduced by this simple rescaling can propagate to the PM-MOD estimates and can be responsible for some ET patterns differing from the official use of the Mu et al. (2011)algorithm for the MODIS ET product.

2.2 Merging technique

2.2.1 Tower weighting

The weights in a merging scheme are typically based on an estimation of some measure of product uncertainty. New description of weighted strategy Here the idea is to estimate the weights proportionally to the agreement between the variations of each ET product and the tower measurements. In order to do so, we propose the following merging scheme:

1. At each tower location, both the different ET products and the tower observations are decomposed into a time series of anomalies and a seasonal climatology as follows:

$$E_m = Ea_m + Ec_m \tag{1}$$

where E_m is the GLEAM (G), PT-JPL (P), PM-MOD (M), and tower observations (O) ET, Ea_m their respective anomalies, and Em_c their respective seasonal climatologies. For the ET products, they are obtained by calculating their respective multi-year (2002–2007) daily

averages. Given the relatively short period, they are further smoothed by applying a 30-day moving average filter. For the towers however, the climatology is estimated over all available site years (even if outside the 2002-2007 period) in order to estimate a climatology that is as robust as possible (note that the obtained climatologies are also further smoothed using the same moving average filter).

2. The product anomalies are weighted as follows:

$$Ea_{WA} = \mathbf{w}^T \mathbf{Ea} \tag{2}$$

where Ea_{WA} is the weighted anomaly, $\mathbf{Ea} = [Ea_G, Ea_P, Ea_M]^T$ is the anomaly vector, and $\mathbf{w} = [w_G, w_P, w_M]^T$ is the weight vector calculated as $\mathbf{w} = (\mathbf{1}^T\mathbf{C1})^{-1}\mathbf{1}^T\mathbf{C}^{-1}$, with $\mathbf{1} = [1,1,1]^T$ and $\mathbf{C}$ the 3x3 error covariance matrix of the differences $Ea_m - Ea_o$ for each product. We expect the errors to have a seasonal dependence. Hence, in order to estimate the temporal evolution of the weights, they are calculated using a moving window, where the error-covariance at a certain point in time is calculated using all available ET estimates within the time window. The choice of window length is subjective: shorter time windows produce more dynamic weights, but their values are likely to be noisier given the smaller number of samples available to estimate the time series variability. A number of 30 days before and after each calendar day was found to provide a good compromise between the smoothness of weights and the number of samples required, so a 61-day running window is used to calculate the daily weights.

3. The merged product is finally calculated by adding the weighted anomalies to the average of the 3 products climatology:

$$E_{WA} = Ea_{WA} + 1/3 \sum_{m=G,P,M} Ec_m \tag{3}$$

where E_{WA} is the weigthed average merged product (WA-merger). Note that the sum of the weights equals one, and that for equally uncertain anomalies the weight vector becomes $[1/3, 1/3, 1/3]^T$. In that case the weighted product corresponds to the simple average (SA-merger) of the individual products.

2.2.2 Weights extrapolation

In order to produce a global weighted product, an extrapolation of the weights from the tower space (i.e., the 84 cells where the towers are located, see Section 3.2) to the entire continental land is needed. The approach chosen to predict the weights outside the tower space is to non-linearly regress the weights based on the main ET model inputs and model ET estimates. For the regression, we use a single neural network (NN) modelling the annual statistical relationship between the weights and their predictors. NNs are broadly used given their capability to approximate non-linear functions,

and are in principle a suitable tool to extrapolate the tower weights. Here it is is used to model the statistical distribution of the weights. However, given that the this error distribution does not only depend on the variables used as predictors in the NN approach, the weights can never be perfectly predicted.

Improving the NN description A standard multi-layer perceptrons with a 11 inputs first layer, one hidden layer with 30 neurons and sigmoidal activation functions, and one output layer with 3 neurons and linear activation functions, is used for the regression. Inputs to the NN are the GLEAM, PT-JPL, and PM-MOD ET together with the surface net radiation, the near-surface air temperature, the relative humidity, the soil moisture, the vegetation optical depth, and the project LAI and FAPAR (see Section 3.1). The outputs to be predicted by the NN are the GLEAM, PT-JPL, and PM-MOD weights. The NN initial weights are randomly initialized by the Nguyen-Widrow algorithm (Nguyen and Widrow, 1990), and the final weights assigned by a Marquardt-Levenberg backpropagation algorithm (Hagan and Menhaj, 1994) minimizing a standard sum of square errors (Bishop, 1995b). Note that given the statistical nature of the prediction, the sum of weights can slightly differ from the expected value of one. To assure the sum equalling one, the NN predicted weights are normalized by their sum.

The objective of any NN is to model the general distribution of the data, not the very specific features of the training dataset. The existence of these specific features is unavoidable, as any training dataset is always limited in terms of being a sample of the true distribution. Modelling the specific features is often referred to as "over-fitting". To prevent the latter standard techniques such as early stopping are applied (Bishop, 1995a). In practice this involves monitoring the evolution of the NN error function for an independent validation data set, here constructed by randomly sampling 20% of the original training data set. While this error decreases at the beginning of the training, there is a moment when starts to increase again. This is taken as an indication of the NN starting to over-fit, and the training is haltered.

Preventing over-fitting only assures the right NN model complexity for the conditions sampled in the training data set. In this particular case the limited spatial coverage of the tower stations suggest a poor sampling of the global conditions (see Section 3.2), and further tests are required to see the NN capacity to extrapolate to un-sampled conditions. For this, we will apply out-sample techniques where one tower station is removed from the training data set, followed by assessing the NN performance at the removed station. If the performance is poor, this strongly suggests that the training data set is not robust enough to represent conditions not sampled within this training data set distribution. Note that for the early-stopping technique training and validation subsets contain data from the same stations. So, if the out-sample technique is also applied, the data from the removed station is no longer part of the training nor validation subsets during the cross-validation.

Note that as tower measurements were masked for rainy intervals (see Section 3.2), the interception loss of the modelled ET is not evaluated. Therefore, only the sum of soil evaporation and

transpiration is compared with the tower data and weighted. To derive the total ET merged product, an estimate of interception loss should also be provided, either by (1) assuming that GLEAM, PT-JPL, and PM-MOD interception loss are equally uncertain and adding their average to the weighted soil evaporation and transpiration, or; (2) by adding just one of the individual model interception losses, if there are reasons to believe that the selected one is less uncertain. Here we adopt the first approach, so the total ET product is the sum of the weighted soil evaporation and transpiration, together with the inter-product interception loss.

2.3 Metrics

Agreement with the towers ET is analyzed by calculating the Pearson correlation coefficient (R), the Mean Square Difference (MSD), and the Root Mean Square Difference (RMSD) according to the expressions:

$$\mathrm{R} = \frac{N\sum_{i=1}^{N} P_i O_i - \sum_{i=1}^{N} P_i \sum_{i=1}^{N} O_i}{\sqrt{[N\sum_{i=1}^{N} {P_i}^2 - (\sum_{i=1}^{N} P_i)^2]}\sqrt{N\sum_{i=1}^{N} {O_i}^2 - (\sum_{i=1}^{N} O_i)^2}} \tag{4}$$

$$\mathrm{MSD} = [\frac{1}{N}\sum_{i=1}^{N}(P_i - O_i)^2] = \mathrm{RMSD}^2 \tag{5}$$

where P and O are the model-derived and observed (or a second model-derived) variate, and N is the number of cases. The MSD can be decomposed into a random (MSD_r) and systematic (MSD_s) component following Willmott (1982) by using the expressions:

$$\mathrm{MSD}_r = \frac{1}{N}\sum_{i=1}^{N}(\hat{P}_i - O_i)^2 = {\mathrm{RMSD}_r}^2 \tag{6}$$

$$\mathrm{MSD}_s = \frac{1}{N}\sum_{i=1}^{N}(P_i - \hat{P}_i)^2 = {\mathrm{RMSD}_s}^2 \tag{7}$$

where $\hat{P}_i = a + bO_i$ is the linear least squares regression of P onto O, being a and b the regression intercept and slope, respectively. Notice that MSD = MSD_r + MSD_s.

Statistics are calculated for the complete study period, or separately for the boreal winter (DJF), spring (MAM), summer (JJA), and autumn (SON). For the correlations, statistical significance is tested by calculating 95% confidence intervals. For the correlation differences, a Fisher Z-transformation is applied to the correlations, and a Student t-test at a 5% significance level used to test the significance of the difference. The autocorrelation of the daily time series is taken into account by reducing the degrees of freedom using an effective sampling size (De Lannoy and Reichle, 2016; Lievens et al., 2017).

3 Data

3.1 Model inputs

The GLEAM, PT-JPL, and PM-MOD required global inputs remain unchanged with respect to Michel et al. (2016) and Miralles et al. (2016), apart from the precipitation product, and are applied at the same resolution of 0.25^o. Common inputs to the models are the surface net radiation, coming from the NASA and GEWEX Surface Radiation Budget (SRB, Release 3.1 Stackhouse et al., 2004), and the near-surface air temperature, sourced from the ERA-Interim atmospheric reanalysis (Dee et al., 2011). PT-JPL and PM-MOD also require near-surface air humidity, also derived from ERA-Interim, and the vegetation products discussed in Sections 2.1.2 and 2.1.3. On the other hand, GLEAM requires precipitation, coming from the Multi-Source Weighted-Ensemble Precipitation (MSWEP) version 1 product (Beck et al., 2017), soil moisture and vegetation optical depth from the European Space Agency (ESA) Climate Change Initiative (CCI) Soil Moisture v2.3 product (Liu et al., 2011b, a), and information on snow water equivalents, from the ESA GlobSnow product for the Northern Hemisphere (Takala et al., 2011), and from the National Snow and Ice Data Center (NSIDC) in snow-covered regions of the Southern Hemisphere (Kelly et al., 2003).

Table 1. List of the FLUXNET sites used in this study together with their FLUXNET code (ID), IGBP land cover (LC) and official reference or principal investigator (PI). The CA-NS1-7 refers to seven stations closely located and run by the same group.

ID	LC	Reference/PI	ID	LC	Reference/PI	ID	LC	Reference/PI
AT-Neu	GRA	George Wohlfahrt	AU-How	SAV	Jason Beringer	BE-Bra	MF	Ivan Janssens
BE-Bra	MF	Ivan Janssens	BE-Lon	CRO	Moureaux et al. (2006)	BE-Vie	MF	Aubinet et al. (2001)
BR-Sa3	EBF	Steininger (2004)	CA-Gro	MF	McCaughey et al. (2006)	CA-Man	ENF	Dunn et al. (2007)
CA-NS1-7	ENF	B.Lamberty et al. (2004)	CA-Oas	MF	Bond-Lamberty et al. (2004)	CA-Obs	ENF	Bond-Lamberty et al. (2004)
CA-Qfo	ENF	Bergeron et al. (2007)	CA-SF1	ENF	Coursolle et al. (2012)	CA-SF2	MF	Amiro et al. (2006)
CH-Dav	ENF	Lukas Hoertnagl	CH-Fru	GRA	Zeeman et al. (2010)	CH-Oe1	GRA	Christof Ammann
CH-Oe2	CRO	Christof Ammann	CN-Cha	MF	Shijie Han	CN-Dan	GRA	Shi Peili
CN-Din	EBF	Guoyi Zhou	CN-Du2	GRA	Chen Shiping	CN-Ha2	WET	Yingnian Li
CN-HaM	GRA	Kato et al. (2006)	CN-Qia	ENF	Huimin Wang	CZ-BK1	ENF	Marian Pavelka
DE-Geb	CRO	Antje Moffat	DE-Gri	GRA	Christian Bernhofer	DE-Hai	DBF	Knohl et al. (2003)
DE-Kli	CRO	Christian Bernhofer	DE-Tha	ENF	Christian Bernhofer	DE-Lnf	DBF	Alexander Knohl
DK-Sor	DBF	Andreas Ibrom	ES-Lju	CSH	Penelope Serrano	FI-Hyy	ENF	Timo Vesala
FR-Fon	DBF	Bazot et al. (2013)	FR-Gri	CRO	Pierre Cellier	FR-LBr	CRO	Denis Loustau
FR-Pu	MF	Jean-Marc Ourcival	IT-Col	DBF	Giorgio Matteucci	IT-Lav	ENF	Damiano Gianelle
IT-MBo	GRA	Damiano Gianelle	IT-PT1	DBF	Günther Seufert	IT-Ren	ENF	Stefano Minerbi
IT-Ro1	CRO	Nicola Arriga	IT-Ro2	DBF	Nicola Arriga	JP-SMF	CRO	Ayumi Kotani
MY-PSO	EBF	Yoshiko Kosugi	NL-Loo	ENF	Eddy Moors	RU-CHE	OSH	Corradi et al. (2005)
RU-Fyo	ENF	Milyukova et al. (2002)	RU-Ha1	GRA	Dario Papale	US-Wi9	MF	Jiquan Chen
US-ARM	CRO	Fischer et al. (2007)	US-ARb	GRA	Margaret Torn	US-ARc	GRA	Margaret Torn
US-Blo	ENF	Goldstein et al. (2000)	US-Cop	GRA	David Bowling	US-IB2	CRO	Roser Mantamala
US-Goo	GRA	Tilden Meyers	US-Ha1	DBF	Goulden et al. (1996)	US-Los	MF	Ankur Desai
US-Ivo	WET	McEwing et al. (2015)	US-MMS	DBF	Schmid et al. (2000)	US-Me2	ENF	Campbell and Law (2005)
US-Me3	ENF	Bond-Lamberty et al. (2004)	US-Ne1	CRO	Simbahan et al. (2006)	US-Ne2	CRO	Amos et al. (2005)
US-Ne3	CRO	Verma et al. (2005)	US-Oho	DBF	Noormets et al. (2008)	US-PFa	MF	Richardson et al. (2006)
US-SRM	WSA	Scott et al. (2009)	US-Syv	MF	Ankur Desai	US-Ton	WSA	Chen et al. (2007)
US-Var	GRA	Ma et al. (2007)	US-WCr	DBF	Cook et al. (2004)	US-Wi3	DBF	Jiquan Chen
US-Wi4	MF	Jiquan Chen						

3.2 Tower data

The FLUXNET 2015 synthesis data set (http://fluxnet.fluxdata.org/) is used to obtain point-based measurements of evaporation (referred to as tower ET), and it is processed as in Martens et al. (2016) to retain only high-quality data appropriate to evaluate the evaporation estimates. Starting from the original time resolution (generally 30 minutes or 1 hour), the processing involves: (1) masking measurements using the originally provided quality flags; (2) masking measurements for rainy intervals, only leaving observations if both the global precipitation product and the local measurements (if available) do not indicate precipitation (as eddy-covariance measurements are less reliable during precipitation events), and; (3) aggregating to daily values if more than 75 percent of remaining sub-hourly data exists for a given day. This quality-check yielded 97 stations. This sample was further reduced to 84 by visually inspecting aerial pictures of the tower surroundings and removing stations close to water bodies, or not representative of the overall land cover within the 0.25^o cells of the gridded ET estimates. The geographical locations of the 84 stations, and their location in an air temperature and precipitation space, are plotted in Fig. 1, with the station names, land covers (based on

the International Geosphere-Biosphere Programme (IGBP) classification), and reference or Principal Investigator listed in Table 1. Note that nearly all stations are in Europe and US, with only two stations located in the Southern Hemisphere.

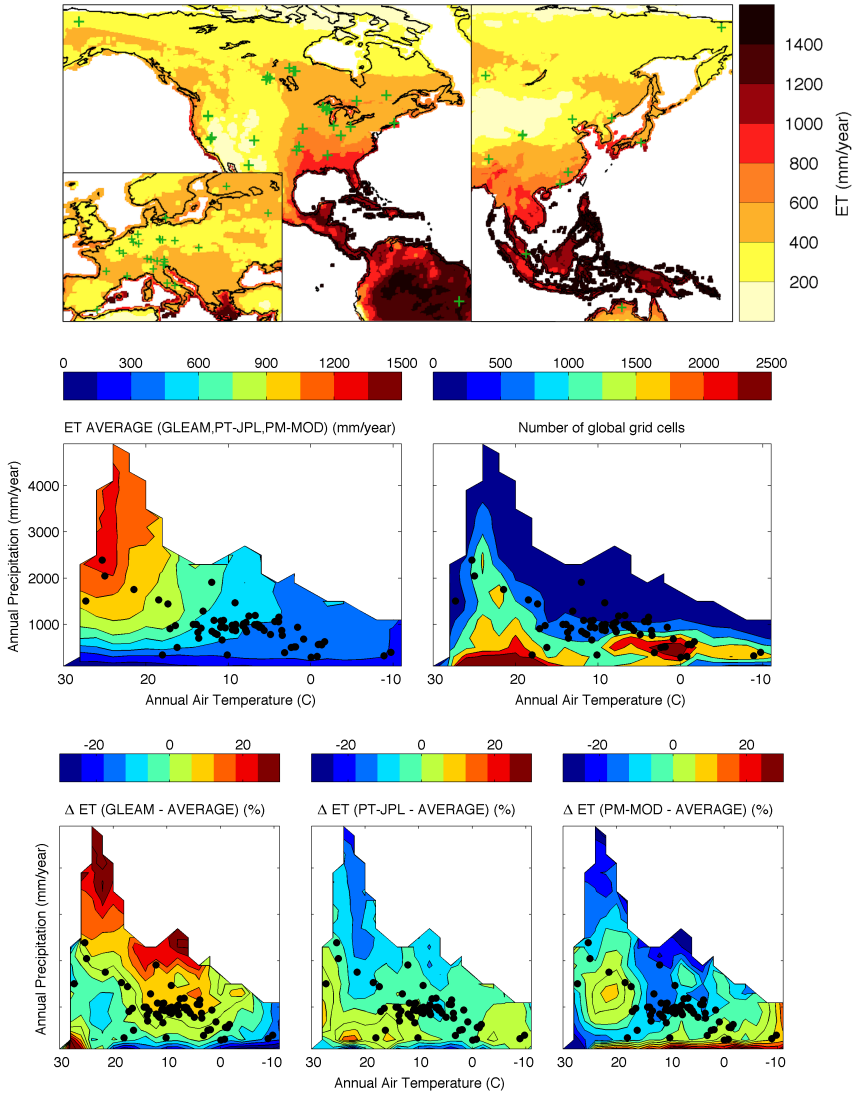

Figure 1. Distribution of tower sites used in the study. Top: geographical location (green crosses) on a map of the multi-annual simple average of the three ET products (GLEAM, PT-JPL, and PM-MOD). Middle: distribution of the averaged multi-annual ET (left), and the number of global grid cells (right), as function of the annual air temperature and precipitation, together with the location of the tower sites in this space (black dots). Bottom: the relative GLEAM (left), PT-JPL (middle), and PM-MOD (right) ET differences normalized by the multi-annual simple average of the three ET products.

Eddy-covariance measurements are subject to errors, both random and systematic, and any merging technique using them as reference is likely to be impacted by those errors. Systematic errors

can arise from instrumental calibration and unmet assumptions about the meteorological conditions, while random errors are typically related to turbulence sampling errors, the assumptions of a constant footprint area, and instrumental limitations (Moncrieff et al., 1996). Estimating these errors is far from simple, and typically requires dedicated experiments (Nordbo et al., 2012; Post et al., 2015; Wang et al., 2015). As such, reporting them is not a widespread practice and error statistics for the individual sites are not commonly available.

The propagation of systematic errors typically results in the lack of energy balance closure observed at many eddy-covariance sites (Wilson et al., 2002; Foken, 2008). Methods to correct the energy unbalance exist, with the Bowen ratio approach (Twine et al., 2000) and the energy balance residual approach (Amiro, 2009) being the most frequently adopted. Corrected fluxes are typically preferred over the original uncorrected observations, but these corrections implies the need for surface radiation and soil heat flux measurements, which are not routinely measured at all stations. At the sites where they are available, the FLUXNET 2015 data set offers a test product containing a corrected version of the heat fluxes based on the Bowen ratio approach, i.e. assuming that the measured Bowen ratio is correct. For the 84 stations selected here, 26 do not have Bowen Ratio Corrected (BRC) fluxes. For the remaining 58 stations, the relative mean difference between the original and BRC latent heat fluxes averaged over all stations is 6.1%, with a maximum value of 16.5%. If the correlation coefficient between original and BRC fluxes is calculated at each station and then averaged over all stations, we obtain 0.96, showing that the original and BRC ET correlate well in time. Also, if the weights of Equation 2 are calculated with the original and BRC fluxes, they display a 0.91 average correlation over all stations and models, with an average RMSD of 0.035. These numbers do not suggest strong differences between both, thus the original (uncorrected) fluxes for all stations are retained for our analyses in order to maximize the number of sites.

Moreover, not all stations cover completely the 2002-2007 period, with 6, 14, 24, 9, and 31 stations reporting 2, 3, 4, 5, and 6 years of data withing the period, respectively. At stations where interannual variability is large the weights may not be representative of the overall climate conditions at the tower if only a relatively short number of years exist. Limiting the study to stations with a relatively large number of years could minimize this drawback, but it would severely reduce the number of towers, so this filtering has not been applied. For instance, if we only derive weights for towers with at least 4 years of data, half of the towers would have been removed. Notice also that due to the masking of the tower data the 61 consecutive daily estimates required to estimate our temporally-varying weights (see Section 2.3) are generally not all available. Therefore, in the case of the tower data we set a minimum threshold of 15 daily values within the 61-day running window for the error to be estimated. Most stations have weights for nearly all days, but in a few stations there are recurrent gaps. A clear example is the tropical BR-Sa3 station, where the frequent rainy episodes complicate the derivation of the weights.

3.3 Ancillary data

Because the substantial mismatch between the size of the model grid cells and the tower footprint is likely to result in representativeness errors, ancillary data sets are required to characterize the spatial homogeneity of the grid cells where the stations are located. Two data sets are considered: the MODIS Land Cover Type product MCD12Q1 at a native resolution of 500 meters, and the Terra MODIS Vegetation Continuous Fields product MOD44B, available at a spatial resolution of 250 meters. A homogeneity index (I_h) is constructed as:

$$I_h = \frac{1}{2} Fgt_{IGBP} + \frac{1}{2}(1 - \mid Fg_{bare} - Ft_{bare} \mid - \mid Fg_{herb} - Ft_{herb} \mid - \mid Fg_{forest} - Ft_{forest} \mid) \quad (8)$$

where Fgt_{IGBP} is the fraction of MCD12Q1 500 meter cells included in the 25 km model grid cell containing the tower and having the same IGBP land cover than the model cell, Ft_{bare}, Ft_{herb} and Ft_{forest} are, respectively, the bare, herbaceous, and forest fractions of the MOD44B 250 meter cell containing the tower, and Fg_{bare}, Fg_{herb} and Fg_{forest} are the same fractions but calculated for the entire 25 km model grid cell where the tower is situated. The first term is the mismatch between the land cover at the tower and at the grid cell level, and the remaining terms are the net mismatch in land cover types across the two resolutions. Ih takes values in the range [0,1], the larger the value the more representative the grid cell is of the landscape of the tower footprint. Finally, to evaluate the merged products, we use river run-off from a compilation of monthly data using different sources, as described in Beck et al. (2015) and annual precipitation estimates from WorldClim Fick and Hijmans (2017) and MSWEP (Beck et al., 2017).

4 Inter-product comparison

The multi-annual GLEAM, PT-JPL, and PM-MOD total ET, together with their absolute and relative differences, are shown in Fig. 2. Differences of the same order can be observed when other products are inter-compared (Jimenez et al., 2011). Given the use of common meteorological forcing (see Section 3.1), the observed differences are mainly introduced by the different approaches to model ET. The disagreement also extends to the the models partitioning of ET into its different components, as shown in Miralles et al. (2016) and (Talsma et al., 2018). We recall here that, as discussed in Section 2.3, only the sum of the soil evaporation and transpiration is compared against tower fluxes.

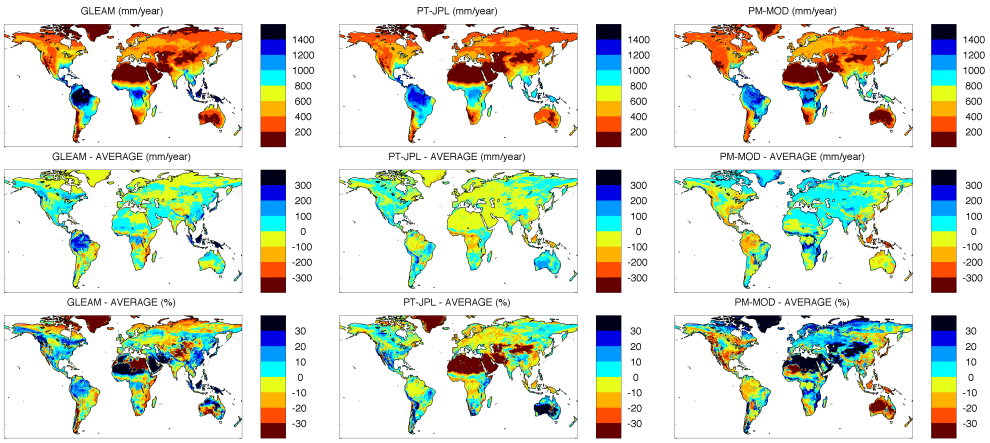

Figure 2. Summary of GLEAM, PT-JPL, and PM-MOD annual ET differences. Top: The GLEAM (left), PT-JPL (middle), and PM-MOD (right) total annual ET in mm/year. Middle: differences between each product and the simple inter-product mean, in mm/year. Bottom: same differences, but normalized by the inter-product mean ET, and expressed as a percentage.

Next, the ET estimates of GLEAM, PT-JPL, and PM-MOD are evaluated at the available tower sites. If we look at the towers spatial distribution in Fig. 1, we can see that there are mostly located in temperate regions. The tropical rain forest and savannas, where the relative ET differences seem larger, are less represented in the selected tower data. Therefore, some regions that would have been relevant to characterize the model ET differences are missing in the evaluation with tower data. Seasonal distributions of ET for three vegetation classes are presented in Fig. 3. The first one includes forest stations (forest), the second one shrublands and savannas (shrub/savanna), and the third one croplands and grasslands (crop/grass). The stations are not evenly distributed within the three groups, with the forest (50 stations) being more represented than the shrubs/savanna and crops/grass (10 and 24, respectively), indicating that summary statistics could be more robust in the case of forests. The surface available energy (Ae) is also plotted. For the models, Ae is the difference between the surface net radiation and the modelled ground flux. For the towers, as the surface net radiation and/or ground flux are not measured at all towers, Ae is given by the sum of the sensible and latent heat fluxes. Clear differences between GLEAM, PT-JPL, PM-MOD and the tower probability distributions are visible. Overall GLEAM and PT-JPL agree better with each other than with PM-MOD, which may be related to the common modelling framework of Priestley-Taylor for GLEAM and PT-JPL, compared with the more different Penman–Monteith approach of PM-MOD.

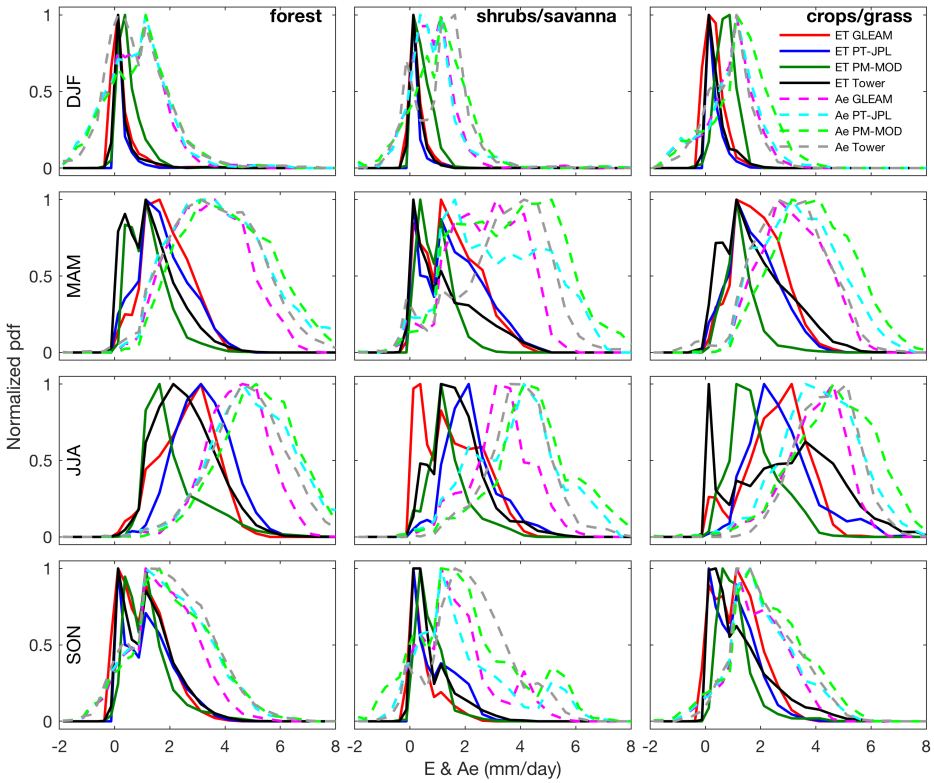

Figure 3. Normalized histograms of ET and available energy (Ae) from GLEAM, PT-JPL, PM-MOD, and the tower observations. The histograms are calculated with the ET values at the tower locations separated first by season and land cover.

An example of good agreement is the forest group in autumn, with the distributions of both ET and Ae being quite similar for the observed and modelled variables. The crops/grass group in summer also shows reasonable agreement between the GLEAM and PT-JPL ET distributions, but larger differences with PM-MOD and the tower ET. In that case, the tower ET shows a clear bimodal distribution, which cannot be replicated any of the models. This may be due to agricultural management practices being poorly captured by the models (e.g., irrigation), but may also reflect the large heterogeneity of croplands and their (a priori) low representativeness of the larger pixel scale. For the shrubs/savanna group during summer, the four ET distributions are quite different, with the Ae distributions also showing differences. For these cases it is difficult to identify whether tower and model ET differences are due to biases in the surface radiation, or discrepancies in the ET formulations.

5 Local merging

5.1 Local weights

Updating the text describing the new weights A summary of daily weight statistics over all the sites belonging to a given land cover group is given in Fig. 4. These weights have been derived based on the differences between the ET product anomalies and the tower ET anomalies as explained in Section 2.2.1. As expected, the simple average product (SA-merger) equally weights all products with a value of 1/3 and is added here as reference. Notice that the weights can take negative values, although the sum of the weights is still one. This happens when the full error covariance matrix has large off-diagonal values reflecting the correlation between the different product errors (e.g. Jones et al., 2008; Hobeichi et al., 2018). This correlation is expected given that the products share some common inputs and model formulations, and it is specially noticeable for GLEAM and PT-JPL. On average, GLEAM has the largest weights and contributes more to the weighted anomalies, but the relative weight of each model is not uniform per season or land cover. For instance, for the forest class PT-JPL is more weighted than GLEAM in winter, while the reverse is true in autumn.

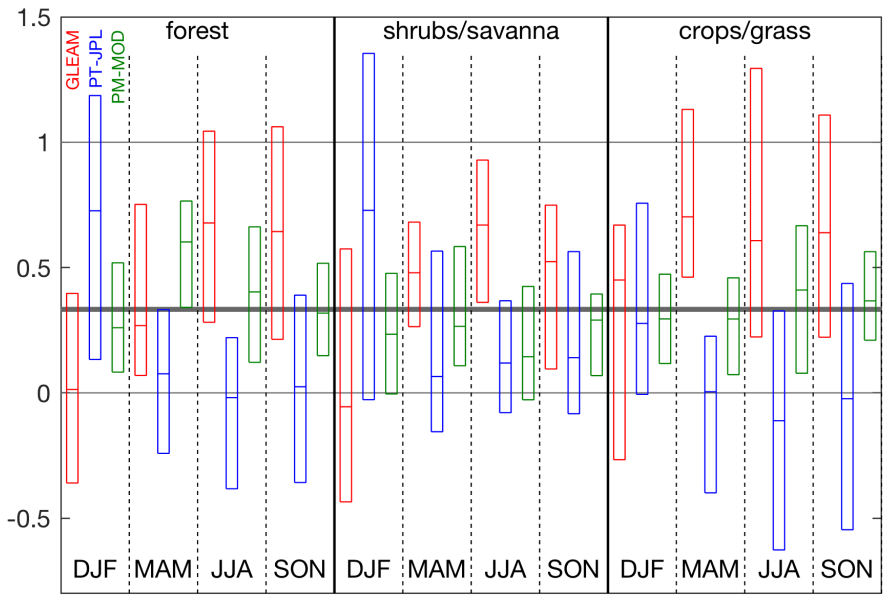

Figure 4. New figure displaying weight statistics Box plots of the GLEAM (red), PT-JPL (blue), and PM-MOD (green) seasonal weights for the three land cover groups. The central mark of the box plots is the median of the group population, the box edges are the 25th ($Q1$) and 75th ($Q3$) percentiles.

An example of the temporal variability of the weights at three towers is given in Fig. 5. At the FR-Pue site, a Mediterranean forest located in France (Rambal et al., 2004), GLEAM starts to be

clearly more weighted for the second part of the year. The correlation between the GLEAM and PT-JPL anomalies is visible in the anti-correlation displayed by the weights. At the US-SRM site, a semi-arid grassland site in southwest of US (Scott et al., 2009), PM-MOD is typically more weighted than GLEAM and PT-JPL in spring, and all weights depart less from the 0-1 range, suggesting more independent errors at this particular station. The last site, the US-Ne1 cropland station situated in North America (Verma et al., 2005), is an example of closer weights for all models for some periods of the year. This happens during the first half of the year. For the second part of the year, the weights change more, with PT-JPL being the most weighted product during some months.

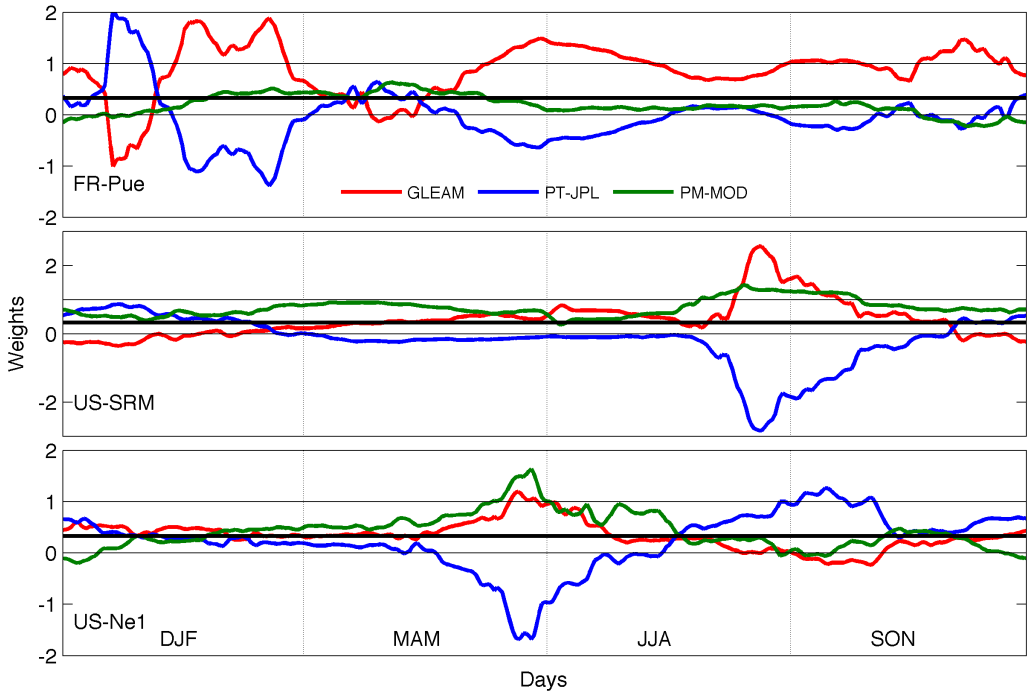

Figure 5. Updating figure with new weights Example of GLEAM (red), PT-JPL (blue), and PM-MOD (green) weights at the FR-Pue (top, forest), US-SRM (middle, grassland), and US-Ne1 (bottom, cropland) stations. The thick black line marks the 1/3 value of the SA-merger weights; the thin black lines mark the 0-1 interval.

5.2 Merged products

Fig. 6 shows – for the same three towers in Fig. 5 – time series of ET from the three products, SA-merger and the weighted average (WA-merger), and the in situ measurement for 2006. Updating text to reflect the new merged products At the FR-Pue site, for this specific year all products disagree with the tower ET for a large part of the year, with PT-JPL and PM-MOD having much larger absolute values overall. Differences between SA-merger and WA-merger are mainly visible in spring and summer, where GLEAM is weighted more strongly, making WA-merger follow the GLEAM estimates more closely. The US-SRM site shows a relatively large ET seasonal variability,

with the ET tightly linked to the precipitation and associated increases in soil moisture (Scott et al., 2009). GLEAM and PT-JPL capture this variability, especially the sudden increase in ET values at the beginning of summer related to the rainfall coming from the North American monsoon. For the first half of summer there are sometimes large differences between SA-merger and WA-merger, with WA-merger correlating better with the tower ET. For the second half, all products fail to replicate the ET increase measured by the tower, and WA-merger and SA-merger are closer to each other as the models anomalies cannot provide information to guide the merging. The US-Ne1 is an irrigated maize-soybean site, where the seasonal cycle of ET is expected to be more pronounced, and accompanied by higher absolute values resulting from irrigation (Verma et al., 2005). The original products have more similar values, not capturing well the ET rise associated with start of the growing season. This may have to do with irrigation not being well captured by any of the models. The closer weights shown in Figure 5 result in closer SA-merger and WA-merger ET.

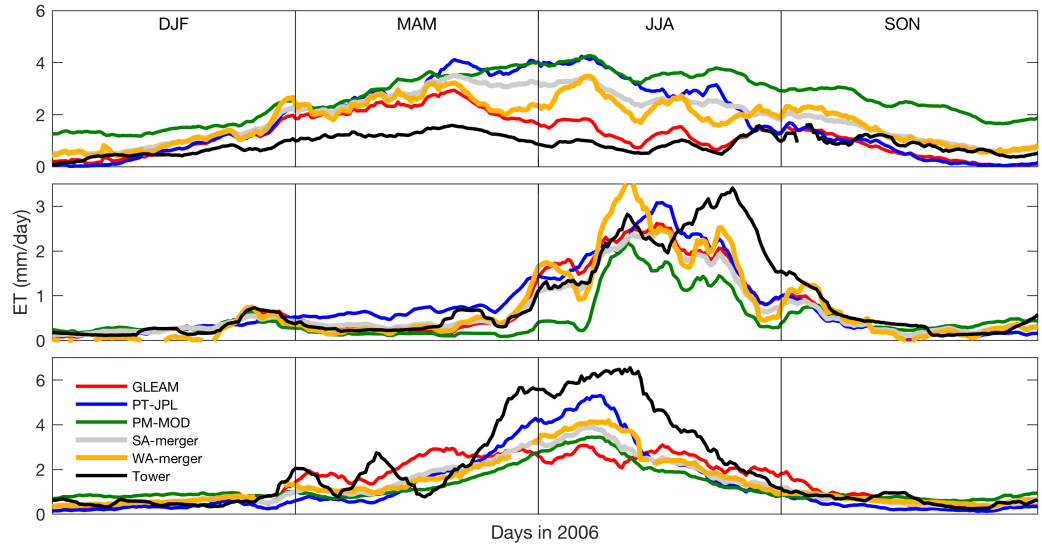

Figure 6. Updating Figure with new weights 2006 time series of the different ET products and the sites shown in Fig. 5: FR-Pue (top), US-SRM (middle), and US-Ne1 (bottom). The daily values are time smoothed using a 10-day moving averaged window to better display the more persistent temporal features.

Updating text to discuss changes in the figure The performance of the individual and merged products across the different stations is summarized in Fig. 7 by plotting seasonal averaged correlations and RMSDs for the three land cover classes. The statistics are presented for the ET anomalies, and for the absolute values. For both, in 10 out of the 12 cases presented (4 seasons x 3 land covers) the correlation of WA-merger is higher than for SA-merger, indicating that an appropriate characterization of the errors – and derived weights – results in better estimates of the ET. The relative increases in correlation between SA-merger and WA-merger are larger for the ET anomalies, but still occur for the ET absolute values. This highlights that when the weighted anomalies are added to the multi-

product climatology, the resulting product combination still overcomes the simple average. Note that the lowest correlations occur in winter time, reflecting the low values and low intra-seasonal variability in this period, while the largest correlations are observed in spring and autumn where vegetation greening and browning typically results in larger ET variability. Note also that correlations are not significant for some stations and periods; non-significant correlations are typically found in wintertime.

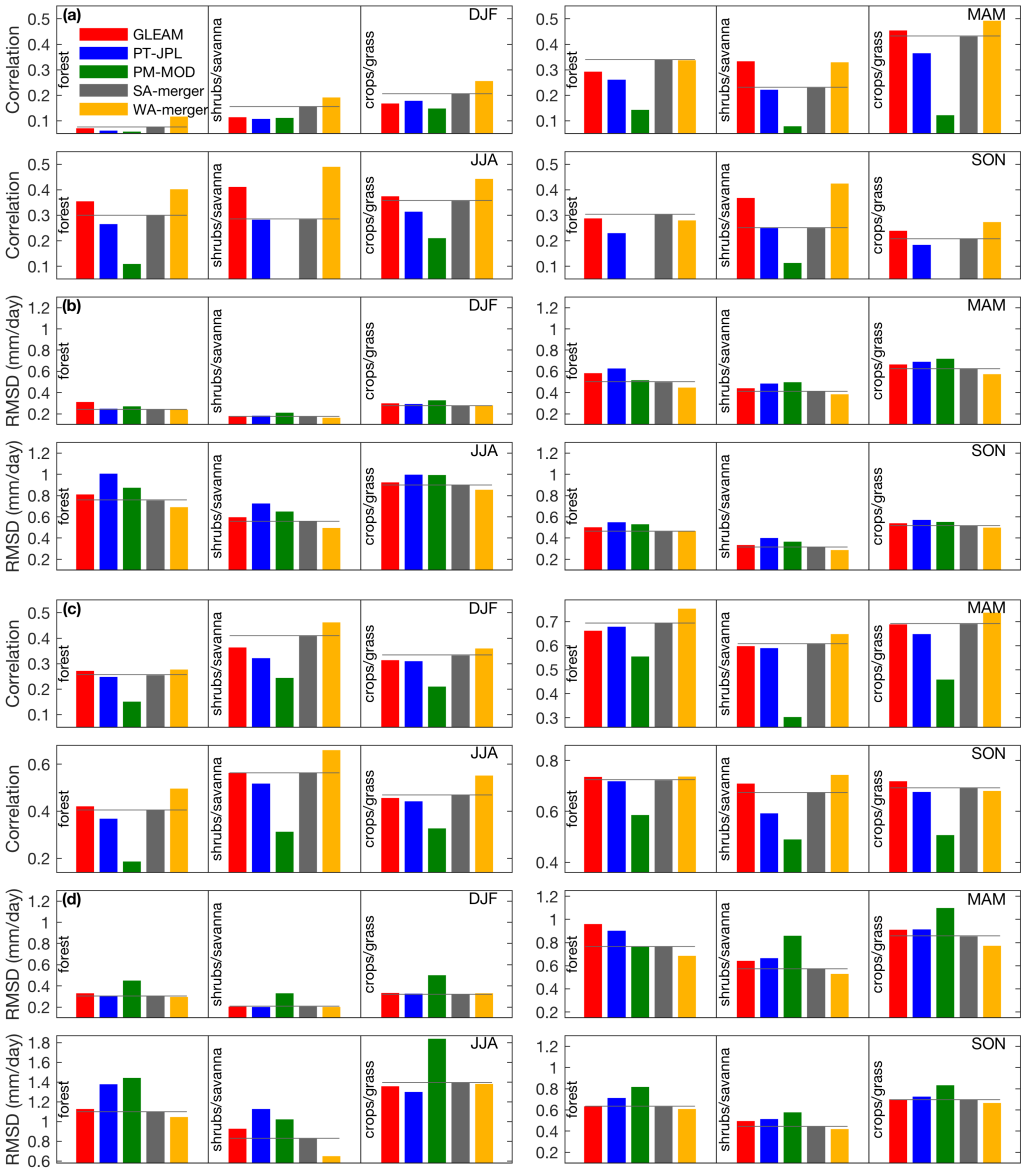

Figure 7. Updating Figure with new SA-merger and WA-merger Season and land-cover averaged ET correlations and RMSD of the tower and the different products (a and b for the ET anomalies, c and d for the ET absolute values). To highlight differences with SA-merger, a grey line has been added to its bar. Note that the axes are not identical, but they cover similar ranges (0.5 for the correlation, 1.2 mm/day for the RMSD).

Concerning the RMSDs, they are slightly lower for WA-merger for all seasons except for winter months. As SA-merger and WA-merger share their climatology (see Section 2.2.1), large differences between both are not expected. This means that the biases between the merged products and the tower ET are preserved for both mergers, indicating that most of the differences in RMSD is coming from changes that are also reflected in the correlations.

6 Global merging

6.1 Global weights

Updating text to discuss the extrapolation of the new weights The local weights at the 84 stations have been extrapolated by the NN as described in Section 2.2.2. The seasonal averages of the weights are presented in Fig. 8. Overall, the spatial patterns of the extrapolated weights for each product do not change substantially across the seasons. Some exceptions are Europe and Northern Asia for GLEAM and PT-JPL. The PM-MOD weights are mostly positive, apart from forested areas in the tropics and some dry areas in Asia and Australia, and are more confined than GLEAM and PT-JPL to the 0-1 interval, indicating smaller error correlation with the other products. For GLEAM and PT-JPL, the weights are a mixture of positive and negative values, and a clear anti-correlation of the weights is visible, i.e., positive GLEAM weights correspond to negative PT-JPL weights, and vice versa, similar to the pattern observed in the local weights for some periods.

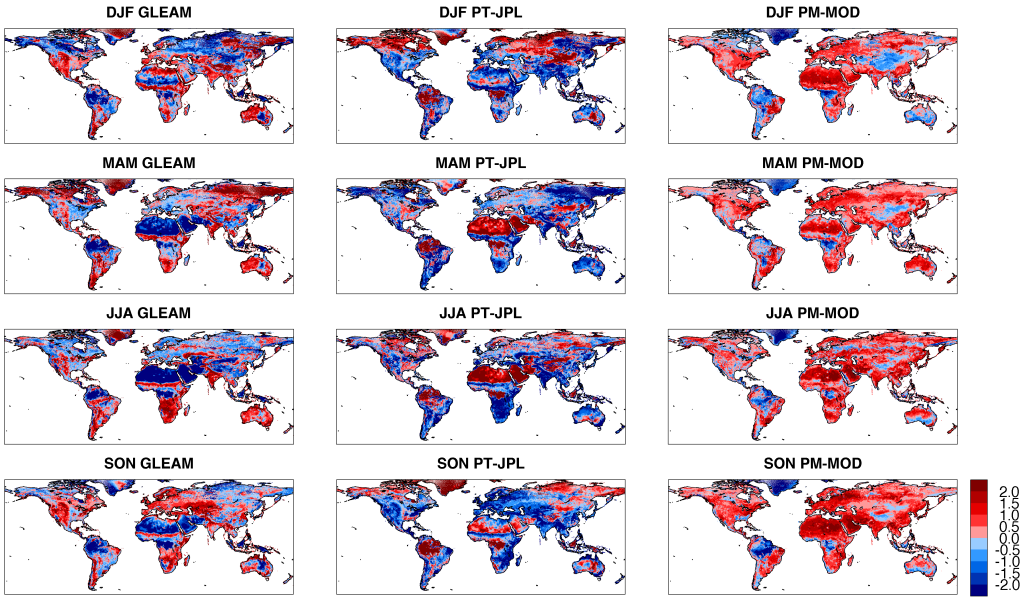

Figure 8. Updating Figure with new weights Seasonally averaged global weights for GLEAM (left), PT-JPL (middle), and PM-MOD (right). Red (blue) colours indicate positive (negative) weights.

6.2 Merged products

Updating text to discuss the merged products after the new weights The seasonally averaged ET differences between WA-merger and SA-merger, normalized by the seasonal SA-merger, are plotted in Fig. 9. The large differences in (semi-)arid areas or the northern latitudes in winter are related to the very low ET absolute values. For the remaining land, most of the relative differences are within the ±25% range. Overall, there are more negative than positive differences, indicating that the WA-merger results in smaller absolute values. Giving that SA-merger and WA-merger have a common climatology, this suggests that the weighting results in an overall reduction of the anomalies at many regions.

Some geographical structures and seasonal changes are visible in some regions. For instance, in the sub-Saharan transition zone the differences are positive in the first half of the year (WA-merger > SA-merger), but negative in the second half. Over India the differences are positive in autumn and winter, but negative in spring and summer. In contrast, some regions do not display large seasonal changes. For instance, in most of Europe WA-merger is smaller than SA-merger over all seasons.

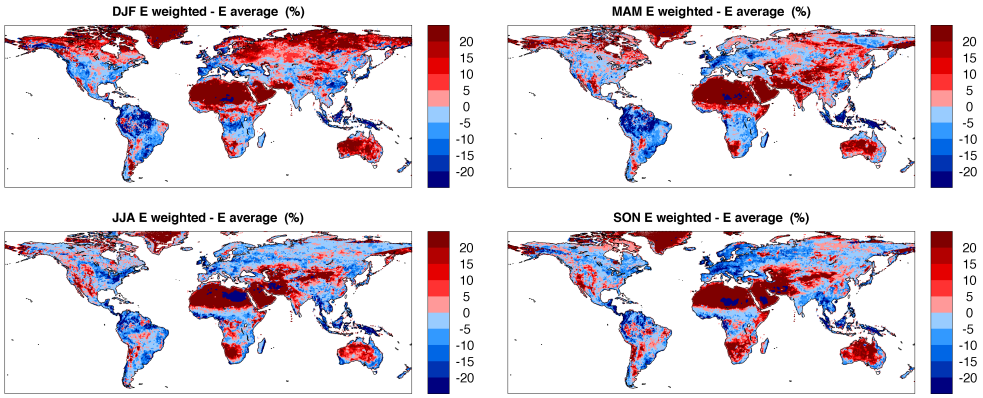

Figure 9. Updating Figure with new products Seasonally averaged normalized ET differences between SA-merger and WA-merger, expressed as a percentage of the seasonally averaged SA-merger ET. Red (blue) colours indicate positive (negative) differences.

7 Discussion

7.1 Tower representativeness

Our inverse error variance weighting is based on the differences between the model and tower ET anomalies. However, it is expected that part of the difference between in situ measurements of ET and model estimates respond to the mismatch in spatial resolution (tower footprint versus model cell). The RMSD of SA-merger against the towers ET, normalized by the mean annual tower ET, is displayed in Fig. 10 for all the available stations, together with the station I_h described in Section 2.3. The towers are sorted from maximum to minimum I_h, i.e., starting by the towers better representing the grid cells where they fall. Nonetheless, low and high normalized RMSDs can occur at stations with comparable I_h, indicating that spatial heterogeneity is only one of the contributing factors to the ET differences. In fact, if the RMSD is linearly regressed on the I_h, the slope of the fit is close to zero, as shown in Fig. 10. Also for the separate products (GLEAM, PT-JPL, and PM-MOD) and WA-merger, no significant correlation between their RMSD against in situ measurements and Ih was found (results not shown). This indicates that for the calculated Ih, and the selected sample of ET products and stations, the error related to the inconsistencies between the tower footprint, and the model pixels does not dominate the total error budget.

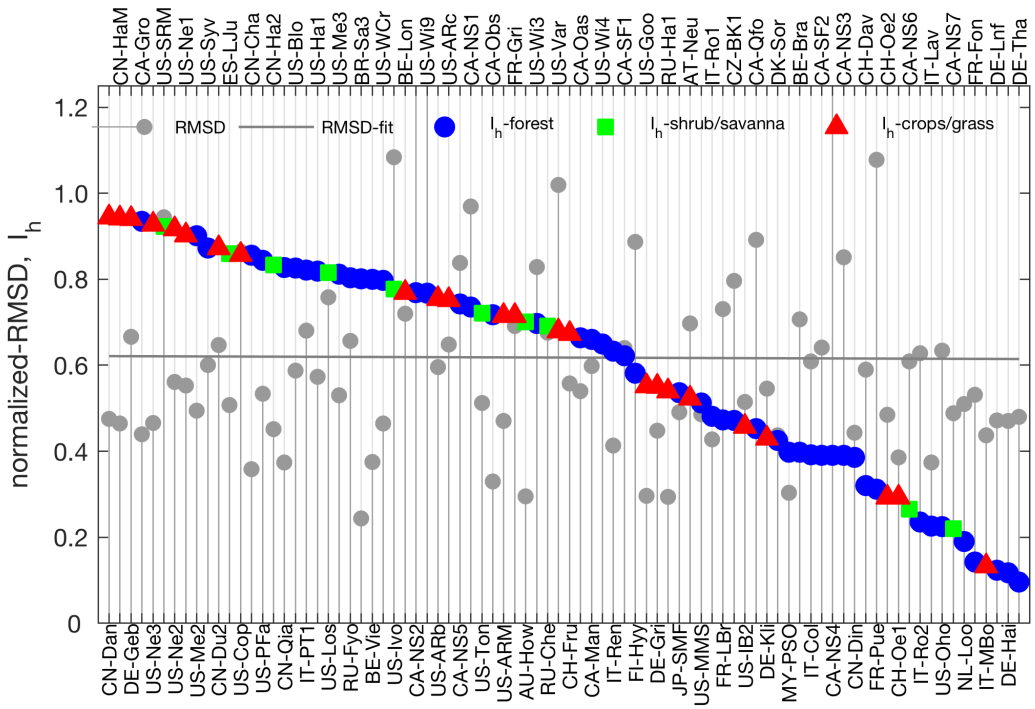

Figure 10. Homogeneity index (I_h) and RMSD of SA-merger and the towers ET. The I_h is plotted as closed circles in blue for forest stations, green for shrubs/savanna, and red for crops/grass, while the RMSD, normalized by the mean annual tower ET, is plotted in grey. A linear fit to the normalized RMSD is given by the grey line. The towers are sorted from maximum to minimum I_h, with the tower names given at the bottom and top of the figure.

7.2 Inverse error variance weighting

Updating text to discuss changes in the random component of the error

The objective on an inverse error-variance weighting is to find the estimate that minimizes the variance of the random error (Rodgers, 2000). As such, the merging only results in the optimal weights if applied over an ensemble of unbiased estimates. Strictly speaking, this requires removing the bias between the model ensemble and in the situ observations prior to the merging, which is not the case here (see Equations 1 to 3). The objective here was to correct the product anomalies towards the tower anomalies, but not to correct the original estimates toward the tower in absolute terms. On the one hand the tower observations have their own systematic errors, as discussed in Section 3.2. On the other hand, debiasing toward the tower ET would require a global correction of the gridded products towards a global tower climatology. If the ultimate objective is to reproduce the tower fluxes, other approaches like regressing the tower ET on either the ET products (Yao et al., 2017) or the ET explanatory drivers (Jung et al., 2011) may appear more straightforward and be possibly more appropriate.

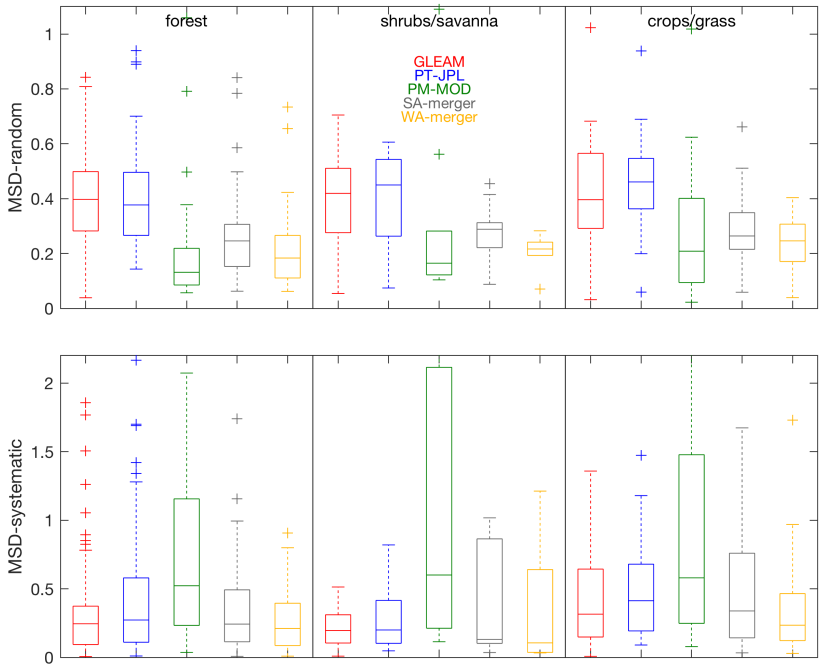

Figure 11. Replacing the components ratio with plots of the individual error components For the three land cover groups the random (top) and systematic (bottom) MSD between the tower ET and GLEAM (red), PT-JPL (blue), PM-MOD (green), SA-merger (grey) and WA-merger (yellow). The central mark of the box plots is the median of the group population, the box edges are the 25th ($Q1$) and 75th ($Q3$) percentiles, the whiskers extend to $Q3 + 1.5(Q3 - Q1)$ and $Q1 - 1.5(Q3 - Q1)$, and values outside the whisker are plotted individually.

Nevertheless, even if optimality in the sense of minimizing the error variance of the WA-merger cannot be assured, weighting the anomalies should result in a decrease of the random error. This is shown in Fig. 11, where box plots of the random (MSD_r) and systematic (MSD_s) components of the difference between the products and the tower observations are displayed (see Equations 6 and 7). From the original products, GLEAM and PT-JPL have comparative error components, while PM-MOD is more distinctive, having smaller MSD_r and larger MSD_s. The latter likely relates to the tendency of the PM-MOD to underestimate ET and its variance (Michel et al., 2016; Miralles et al., 2016). Comparing WA-merger to SA-merger, the reduction of the MSD_r for WA-merger is indicative of the merging being effective in this regard. There is also a slight reduction in the MSD_s, with WA-merger having the smallest median error of all products.

7.3 Weights extrapolation

The number of stations used in this merging exercise is certainly limited in terms of covering different biomes and climatic conditions. Hence, the ability to represent the full distribution of ET across time, space, and biomes is questionable. Updating text to discuss new weights and the test where

a number of days is removed from the training data set. This is verified here by out-sampling the NN training data set in two different ways. In the first test all stations are included in the tower data set, i.e., the standard configuration used to produce the global WA-merger. Before training the NN, 15% of the days at each station are randomly masked from the training data set, and the prediction statistics are derived over this independent subset. In the second test, the station where the prediction will be checked is entirely removed from the training data set, i.e, the weights for that station are derived using a NN that did not include that station in the training phase (i.e. leave-one out cross validation).

A box plot summarizing the correlation and RMSD between the station weights and the weights predicted by the NN for these two tests is presented in Fig. 12. The results clearly show that the correlation and RMSDs between the predicted and the original weights at the stations degrades notably when stations are fully removed from the training data set. This implies that the global extrapolation of the weights will be quite uncertain for conditions not sampled in the available tower data set. For some stations, the out-sampling from the training data set does not have a large effect, because the mapping between the predictors and ET can still be approximated from the relationship presented by other stations. This is for instance the case for the Canadian forest stations CA-NS1-7 (results not shown). However, for other stations, the statistics are good when predicted with the standard data set, but poor with the one-station-removed data set, indicating that the particular conditions of those stations are not well represented in the out-sampled data set. This happens for stations such as US-Wi4 (forest with a snowy winter and warm humid summer) and CN-Dan (grasslands with a polar tundra climate). Finally, there are also stations where statistics are rather poor in both tests, indicating that a link between the model inputs and the related output error could not be established. This is the case for stations such as IT-Col (deciduous broadleaf forest with temperate climate) or MY-Pso (tropical forest). This guaranties that the extrapolation of weights to areas with similar conditions will be very uncertain, even if those conditions were represented in the tower data set.

An additional test to check the representativeness of the tower data set is conducted by globally extrapolating the weights with each of the previous 84 NNs trained without one station, and then checking the variability of the predicted weights. For the conditions well represented in the training data set, it is expected that removing one station will only result in slight changes in the extrapolated weights. However, for regions that are poorly represented, a slightly different data set is likely to result in substantially different weights. This is illustrated in Fig. 13, where a weight variability index is displayed. The index is calculated by: (1) estimating for each global cell the annual standard deviation of the GLEAM, PT-JPL, and PM-MOD weights, normalized by the sum of the absolute annual model weights, and; (2) averaging this standard deviation over the three models. To facilitate its display in Fig. 13, it has been scaled to span the range 0-1. Overall the variability is larger in the Southern Hemisphere than in the Northern Hemisphere, which is expected given that all stations but two are situated in the Northern Hemisphere. The smallest variability in the weights coincides

with the regions where the database is more representative, namely US, Central Europe, and some parts of Asia, suggesting a bias in the tower data set linked to the specific location of the towers selected. The variability in tropical regions, where only 3 stations are part of the database, is in general larger than for the previous regions. The largest variability occurs over the very dry regions, a regime poorly represented in the tower data set as shown in Fig. 1. While a poor extrapolation of weights is not critical over very dry regions, given their low ET values, uncertain weights over the very humid regions is more of a concern due to their typically large ET values and their significance for the global mean ET.

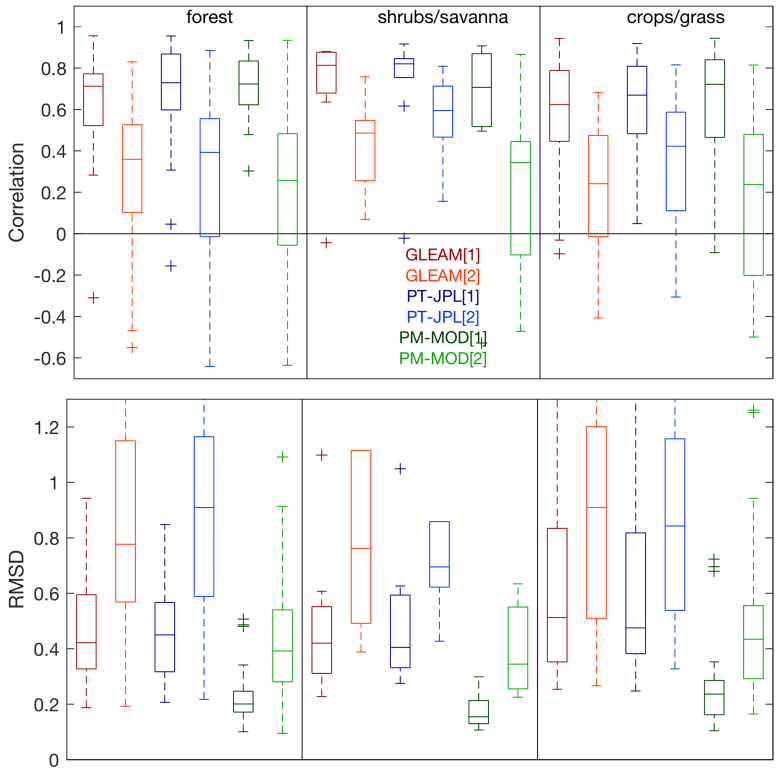

Figure 12. Updating Figure with newly extrapolated weights Box plot showing for the three land cover groups the correlation (top) and RMSD (bottom) between the station local weights and the weights predicted by the NN for the two tests presented in Section 7.3 (first test labelled as [1] in legend, dark colours, second test as [2], light colours). See the text for details.

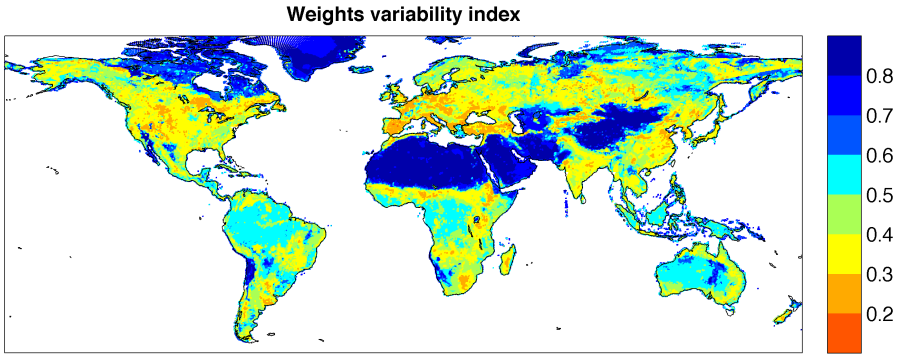

Figure 13. Relative annual variability of the global weights extrapolated by 84 different NNs. Smaller (larger) values indicate lower (higher) variability. See the text for details.

7.4 Merged products evaluation

The evaluation of ET products is typically conducted by comparing the estimates to point scale tower fluxes. Alternatively, water balance calculations at larger spatial scales – such as catchment scales - where ET is estimated as the residual of precipitation (P) and river run-off (Q) are often used as well. As the towers are used to derive the merge products, the alternative for an independent assessment of the merged products is to conduct such catchments mass balance analyses (e.g. Vinukollu et al., 2011; Miralles et al., 2016). This assessment appears a good means to evaluate the long-term mean ET estimates. Nonetheless, as WA-merger and SA-merger share a common mean state, large performance differences are not expected. Note that to retain the independence, the precipitation used in the water balance calculation should not be the one used as forcing in the ET estimates. Here this is an issue for GLEAM and the merged products (as they include GLEAM), but not for PT-JPL and PM-MOD as they do not use precipitation data as input. As such, WordlClim precipitation data is also used in addition to MSWEP in these comparisons (GLEAM was forced with MSWEP, see Section 3.1).

The mass balance of a catchment implies that the space and time integration of P-Q should equal the ET integrated over the same space and time, if one assumes that the the changes in soil water storage within the catchment are small compared with the cumulative volume of ET, P, and Q. The longer the period, the more valid this assumption becomes. Here, the mean 2002-2007 ET estimates from GLEAM, PT-JPL, PM-MOD, and the merged products are calculated per catchment. The basin P-Q estimate is then calculated using the Q and P data described in Section3.3. We only select catchments for which the P-Q data record is available for a minimum of 3 years in the 2002-2007 period, to assure some common period between ET and P-Q. In addition, to reduce noise in the basin-integrated ET estimates, only basins with a catchment area containing at least 3x3 cells of the

25 km resolution gridded estimates are included in the comparison. This results in 685 basins, 75 % of them situated in the Northern Hemisphere (i.e showing a similar geographical bias as the tower data set). Catchments are further divided into three groups of 243, 295, and 147 basins based on the aridity index (AI, basin potential ET over the basin P) taking values in the intervals AI<1, 1<AI<2, and AI>2.

Updating text to reflect changes in the merged products Scatter plots showing the correspondence between P-Q and ET are given in Fig. 14. Linear fits for the three AI classes are plotted, and the correlation, RMSD, and bias (ET minus P-Q) given in the plot. Overall, the statistics of the the water balance comparison using MSWEP or WordlClim as P are close, suggesting that the dependence on MSWEP is not a determining factor in the agreement. From the original products, PM-MOD shows the worst agreement with P-Q. GLEAM agrees better than PT-JPL for the wettest and specially for the driest basins. For the latter, GLEAM shows correlations of 0.93 (based on MSWEP) and 0.88 (based on World-Clim), compared to the respective 0.74 and 0.69 for PT-JPL. However, PT-JPL agrees slightly better than GLEAM for the 1<AI<2, although both show similar correlations. However, PT-JPL agrees slightly better than GLEAM for the 1<AI<2, although both show similar correlations. The SA-merger shows close statistics to GLEAM and PT-JPL, so adding the PM-MOD product neither improves nor degrades the skill to close the catchment water budget. Regarding a comparison between WA-merger and SA-merger, their statistics are very close. Correlations are comparable, and only in terms of RMSD WA-merger ET agrees slightly better with P-Q for the wettest basins (AI>2), with 89/83 (MSWEP/WordlClim) mm/year RMSDs and -64/-46 mm/year biases for the WA-merge product, and 115/107 mm/year RMSDs and -97/-80 mm/year biases for the SA-product. Notice also that for the wettest basins these WA-merger performs better than any individual product.

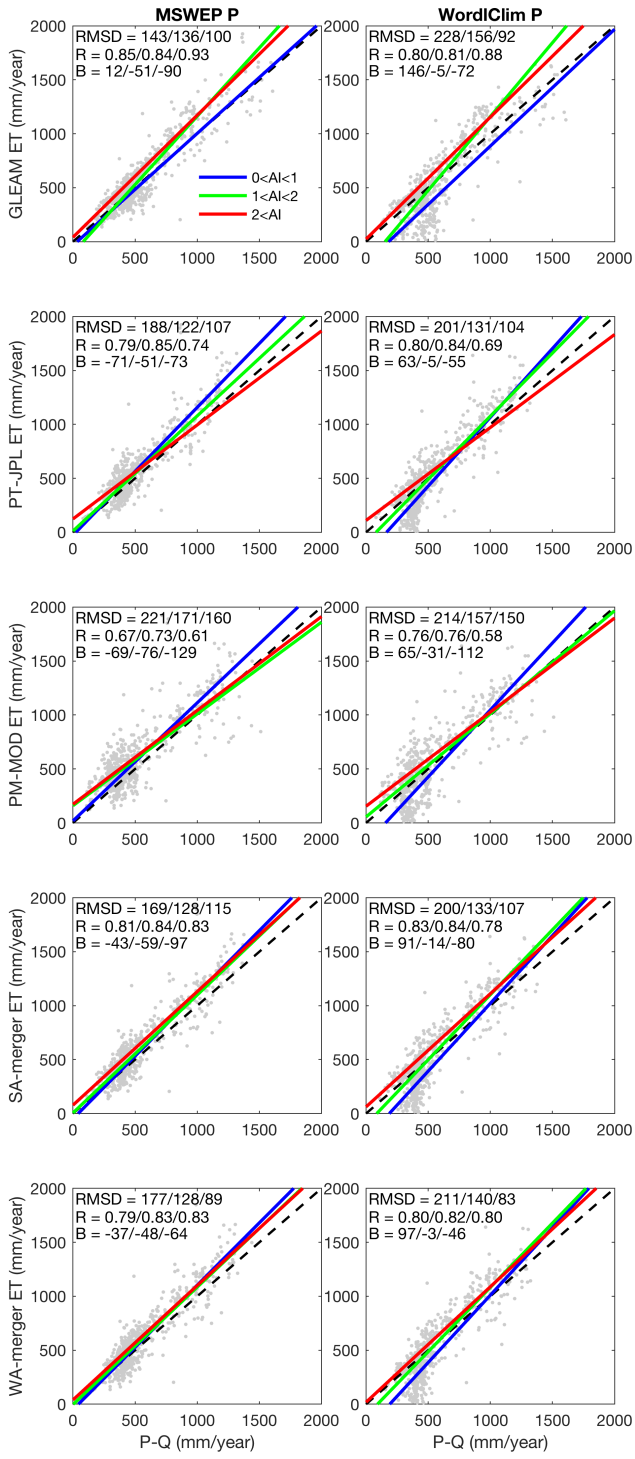

Figure 14. Replacing slope with bias Scatter plots of P-Q and ET from the different products. Linear fits for three AI classes are plotted, together with the correlation, RMSD and bia s (ET - (P-Q)). From left to right, the statistics are given for AI<1 (blue line), 1<AI<2 (green), and AI>2 (red), i.e., from dry to wet basins.

8 Conclusions

Updating to reflect changes in the manuscript

A simple average (SA-merger) and an inverse error variance weighting (WA-merger) of the three global ET products generated during the WACMOS-ET project is presented. During the project, three ET models were forced with common daily inputs at a resolution of 25 km for the period 2002-2007: GLEAM, PT-JPL, and PM-MOD. GLEAM and PT-JPL share a Priestley-Taylor formulation to estimate potential evaporation, while PM-MOD uses a more different modeling approach of potential evaporation based on a Penman-Monteith formulation, but a very similar evaporative stress and radiation partitioning formulation to the one by PT-JPL. In WA-merger, the weights were estimated using the error-variance of the individual product anomalies, with the error defined as the difference between tower-based ET anomalies and modeled ET anomalies for non-rainy conditions. Then the final data set was reconstructed by adding the weighted anomalies to the mean seasonal climatology of the products. A similar approach was followed to generate SA-merger, but in this case giving equal weights to the anomalies of all three products. Finally, the potential to extrapolate these locally-estimated weights to the global scale based on a neural network approach has been explored. Given the described framework, the intent here is to evaluate the potential of blending these data sets to yield anomalies of ET that better represent those measured by the global network of eddy-covariance towers. We note that capturing anomalies in ET is crucial for applications such as drought monitoring or irrigation planning.

The resulting local weights showed seasonal patterns and negative values at many stations. This was to a large extent related to correlation in the errors of the anomalies of GLEAM and PT-JPL. Nonetheless, seasonal correlations between WA-merger and the tower ET are overall higher than for the individual products and SA-merger. This is mostly attributed to a successful reduction in the random error. Meanwhile, the globally extrapolated weights showed seasonal and regional variability, with these patterns resulted in seasonal differences between the global SA-merger and WA-merger of up to 25% in a large number of regions. However, the limited global coverage of the tower stations, mostly located in the Northern Hemisphere temperate regions, casted doubts on the ability of the NN prediction scheme to reliably extrapolate the locally-estimated weights. This was apparent when the extrapolation was tested over individual stations with the training data set not including the station under study, and when reproducing the global extrapolation of the weights with the training data set missing one station at a time. Both mergers were also compared with the ET inferred from water balance calculations in different catchments across the globe, and similar correlations and RMSDs were obtained, with only slightly better results for the WA-merger over wet basins.

Several limiting factors for the merging exercise are identified, some of which could be informative for other initiatives aiming to blend ET data sets. A longer study period can give access to more in situ data and extend the in situ data set to less represented regions. This would clearly help the global extrapolation of the weights. In addition, the mismatch between the spatial resolution of the

towers and the products is still an issue, despite the fact that here other error sources were deemed to be more dominant. The impact of the mismatch in spatial resolution is expected to be minimized as ET datasets move towards finer spatial resolutions. Dependency between the ET products can also have an impact on the merged products. In this study the GLEAM, PT-JPL, and PM-MOD products are derived with common data sets for their shared inputs. While this was motivated by the primary objective of WACMOS-ET of studying algorithm differences, this is can become a drawback when aiming to achieve an optimal merger. In that case a lower inter-dependency is expected to be beneficial.

Overall, our study suggests that an inverse error variance scheme combining information from tower observations and ET products has the potential to improve upon the simple mean proposed by several previous efforts (e.g. Mueller et al., 2013). However, care should be taken regarding the dependence of the products to be merged, the tower coverage, the different product errors, the spatial representativeness of the in situ measurements at the products resolution, and the nature of the errors of the ET products. Critical for the success of the merging scheme is the adequate characterization of the uncertainty of the individual products, and finding an effective method to extrapolate the weights from the tower space to the global landscape. The latter seems challenging, and given the difficulties found here, alternatives should be considered. A possibility could be triple collocation (Yilmaz et al., 2012). This technique would require two new global ET data sets independent from the products that need to be merged . This can be demanding, but work in that direction has already started (Khan et al., 2018). An added advantage of this approach will be that the tower observations could then be used as an independent evaluation set, similar to the approach carried out for some other Earth Observation products, such as the soil moisture estimates from the ESA Climate Change Initiative (Gruber et al., 2017). This can be of importance, given the very few existing data sets that can be used to presently evaluate ET estimates.

Acknowledgements. This study was funded by the European Space Agency (ESA) and conducted as part of the project WACMOS-ET-Ensemble (ESRIN Contract No. 4000117355/16/I-NB). D.G.Miralles acknowledges support from the European Research Council (ERC) under grant agreement number 715254 (DRY-2-DRY). J.B. Fischer contributed to this paper at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration. California Institute of Technology. Government sponsorship acknowledged. Support to J.B Fisher was provided by NASA's SUSMAP, THP, and INCA programs, and the ECOSTRESS mission. K. Tu, from the Department of Ecosystem and Conservation Sciences, University of Montana, is acknowledged by providing guidance for the use of the vegetation products in this study. This work used eddy covariance data acquired by the FLUXNET community and in particular by the following networks: AmeriFlux (U.S. Department of Energy, Biological and Environmental Research, Terrestrial Carbon Program (DE-FG02-04ER63917 and DE-FG02-04ER63911)), AfriFlux, AsiaFlux, CarboAfrica, CarboEuropeIP, CarboItaly, CarboMont, ChinaFlux, Fluxnet-Canada (supported by CFCAS, NSERC, BIOCAP, Environment Canada, and NRCan), GreenGrass, KoFlux, LBA, NECC, OzFlux, TCOS-Siberia, USCCC.

Data and logistical support for the station US-Wrc were provided by the US Forest Service Pacific Northwest Research Station. The FLUXNET eddy-covariance data processing and harmonization were carried out by the ICOS Ecosystem Thematic Center, the AmeriFlux Management Project, and the Fluxdata project of FLUXNET, with the support of CDIAC, and the OzFlux, ChinaFlux, and AsiaFlux offices.

References

Aires, F.: Combining Datasets of Satellite-Retrieved Products. Part I: Methodology and Water Budget Closure, Journal of Hydrometeorology, 15, 1677–1691, 2014.

Amiro, B.: Measuring boreal forest evapotranspiration using the energy balance residual, Journal of Hydrology, 366, 112–118, 2009.

Amiro, B., Barr, A., Black, T., Iwashita, H., Kljun, N., Mccaughey, J., Morgenstern, K., Murayama, S., Nesic, Z., and Orchansky, A.: Carbon, energy and water fluxes at mature and disturbed forest sites, Saskatchewan, Canada, Agricultural and Forest Meteorology, 136, 237–251, 2006.

Amos, B., Arkebauer, T. J., and Doran, J. W.: Soil surface fluxes of greenhouse gases in an irrigated maize-based agroecosystem, Soil Science Society of America Journal, 69, 387–395, doi:10.2136/sssaj2005.0387, 2005.

Aubinet, M., Chermanne, B., Vandenhaute, M., Longdoz, B., Yernaux, M., and Laitat, E.: Long term carbon dioxide exchange above a mixed forest in the Belgian Ardennes, Agricultural and Forest Meteorology, 108, 293–315, 2001.

Baldocchi, D., Falge, E., Gu, L., Olson, R., Hollinger, D., Running, S., Anthoni, P., Bernhofer, C., Davis, K., Evans, R., Fuentes, J., Goldstein, A., Katul, G., Law, B., Lee, X., MALHI, Y., Meyers, T., Munger, W., Oechel, W., Paw, K. T., Pilegaard, K., Schmid, H. P., Valentini, R., Verma, S., Vesala, T., Wilson, K., and Wofsy, S.: FLUXNET: A New Tool to Study the Temporal and Spatial Variability of Ecosystem-Scale Carbon Dioxide, Water Vapor, and Energy Flux Densities., Bulletin of the American Meteorological Society, 82, 2415–2434, 2001.

Bazot, S., Barthes, L., Blanot, D., and Fresneau, C.: Distribution of non-structural nitrogen and carbohydrate compounds in mature oak trees in a temperate forest at four key phenological stages, Trees, 27, 1023–1034, 2013.

Beck, H. E., De Roo Journal of, A., and van Dijk, A. I. J. M.: Global maps of streamflow characteristics based on observations from several thousand catchments, journals.ametsoc.org, 16, 1878–1501, 2015.

Beck, H. E., Vergopolan, N., Pan, M., Levizzani, V., van Dijk, A. I. J. M., Weedon, G. P., Brocca, L., Pappenberger, F., Huffman, G. J., and Wood, E. F.: Global-scale evaluation of 22 precipitation datasets using gauge observations and hydrological modeling, Hydrology and Earth System Sciences, 21, 6201–6217, 2017.

Bergeron, O., MARGOLIS, H. A., BLACK, T. A., COURSOLLE, C., Dunn, A. L., BARR, A. G., and Wofsy, S. C.: Comparison of carbon dioxide fluxes over three boreal black spruce forests in Canada, Global Change Biology, 13, 89 – 107, doi:10.1111/j.1365-2486.2006.01281.x, 2007.

Betts, A. K.: Land-surface-atmosphere coupling in observations and models, Journal of Advances in Modeling Earth Systems, 2, 4–18, 2009.

Bishop, M.: Neural networks for pattern recognition, chap. Learning and Generalization, pp. 332–384, Oxford University Press, Inc., New York, 1995a.

Bishop, M.: Neural networks for pattern recognition, chap. Error functions, pp. 194–252, Oxford University Press, Inc., New York, 1995b.

B.Lamberty, B., Wang, C., and Gower, S. T.: Net primary production and net ecosystem production of a boreal black spruce wildfire chronosequence, Global Change Biology, 10, 473 – 487, doi:10.1111/j.1529-8817.2003.0742.x, 2004.

Bond-Lamberty, B., Wang, C. K., and Gower, S. T.: Net primary production and net ecosystem production of a boreal black spruce wildfire chronosequence, Global Change Biology, 10, 473–487, doi:10.1111/j.1529-8817.2003.0742.x, 2004.

Campbell, J. L. and Law, B. E.: Forest soil respiration across three climatically distinct chronosequences in Oregon, Biogeochemistry, 73, 109–125, 2005.

Chen, Q., Gong, P., Baldocchi, D., and Tian, Y. Q.: Estimating basal area and stem volume for individual trees from lidar data, Photogrammetric Engineering and Remote Sensing, 73, 1355–1365, doi:http://dx.doi.org/10.14358/PERS.73.12.1355, 2007.

Cook, B. D., Davis, K. J., Wang, W. G., Desai, A., Berger, B. W., Teclaw, R. M., Martin, J. G., Bolstad, P. V., Bakwin, P. S., Yi, C. X., and Heilman, W.: Carbon exchange and venting anomalies in an upland deciduous forest in northern Wisconsin, USA, Agricultural and Forest Meteorology, 126, 271–295, doi:10.1016/j.agrformet.2004.06.008, 2004.

Corradi, C., Kolle, O., Walter, K., Zimov, S. A., and Schulze, E.-D.: Carbon dioxide and methane exchange of a north-east Siberian tussock tundra, Global Change Biology, pp. 1910–1925, doi:10.1111/j.1365-2486.2005.01023.x, 2005.

Coursolle, C., Margolis, H. A., Giasson, M.-A., Bernier, P.-Y., Amiro, B., Arain, M. A., Barr, A., Black, T. A., GOULDEN, M. L., McCaughey, J., Chen, J., Dunn, A., Grant, R. F., and Lafleur, P.: Influence of stand age on the magnitude and seasonality of carbon fluxes in Canadian forests, Agricultural and Forest Meteorology, 165, 136 – 148, doi:10.1016/j.agrformet.2012.06.011, 2012.

De Lannoy, G. and Reichle, R. H.: Global assimilation of multiangle and multipolarization SMOS brightness temperature observations into the GEOS-5 catchment land surface model for soil moisture . . . , Journal of Hydrometeorology, 17, 669–691, 2016.

Dee, D., Uppala, M., S., Simmons, J., A., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, A., M., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, M., A. C., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, J., A., Haimberger, L., Healy, B., S., Hersbach, H., Hólm, V., E., Isaksen, L., Kallberg, P., Khaler, M., Matricardi, M., McNally, P., A., Monge-Sanz, M., B., Morcrette, J.-J., Park, B.-K., Peubey, C., de Rosnay, P., Tavolato, C., Thapaut, J.-N., and Vitart, F.: The ERA-Interim reanalysis: configuration and performance of the data assimilation system, Q. J. R. Meteorol. Soc., 137, 553–597, doi:10.1002/qj.828, 2011.

Dunn, A. L., Barford, C. C., Wofsy, S. C., Goulden, M. L., and Daube, B. C.: A long-term record of carbon exchange in a boreal black spruce forest: means, responses to interannual variability, and decadal trends, Global Change Biology, 13, 577 – 590, doi:10.1111/j.1365-2486.2006.01221.x, 2007.

Dunn, S. M. and Mackay, R.: Spatial variation in evapotranspiration and the influence of land use on catchment hydrology, Journal of Hydrology, 171, 49–73, 1995.

Fick, S. E. and Hijmans, R. J.: WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas, International Journal of Climatology, 37, 4302–4315, 2017.

Fischer, M. L., Billesbach, D. P., Berry, J. A., Riley, W. J., and Torn, M. S.: Spatiotemporal variations in growing season exchanges of CO2, H2O, and sensible heat in agricultural fields of the Southern Great Plains, Earth Interactions, 11, 1–21, 2007.

Fisher, J. B., Tu, K. P., and Baldocchi, D. D.: Global estimates of the land–atmosphere water flux based on monthly AVHRR and ISLSCP-II data, validated at 16 FLUXNET sites, Remote Sensing of Environment, 112, 901–919, 2008.

Fisher, J. B., Melton, F., Middleton, E., Hain, C., Anderson, M., Allen, R., McCabe, M. F., Hook, S., Baldocchi, D., Townsend, P. A., Kilic, A., Tu, K., Miralles, D. D., Perret, J., Lagouarde, J.-P., Waliser, D., Purdy, A. J., French, A., Schimel, D., Famiglietti, J. S., Stephens, G., and Wood, E. F.: The future of evapotranspiration: Global requirements for ecosystem functioning, carbon and climate feedbacks, agricultural management, and water resources, Water Resources Research, 53, 2618–2626, 2017.

Foken, T.: The energy balance closure problem: an overview, Ecological Applications, 18, 1351–1367, 2008.

Goldstein, A. H., Hultman, N. E., Fracheboud, J. M., Bauer, M. R., Panek, J. A., Xu, M., Qi, Y., Guenther, A. B., and Baugh, W.: Effects of climate variability on the carbon dioxide, water, and sensible heat fluxes above a ponderosa pine plantation in the Sierra Nevada (CA), Agricultural and Forest Meteorology, 101, 113–129, doi:http://dx.doi.org/10.1016/S0168-1923(99)00168-9, 2000.

Goulden, M. L., Munger, J. W., Fan, S. M., Daube, B. C., and Wofsy, S. C.: Measurements of carbon sequestration by long-term eddy covariance: Methods and a critical evaluation of accuracy, Global Change Biology, 2, 169–182, doi:10.1111/j.1365-2486.1996.tb00070.x, 1996.

Gowda, P. H., Chavez, J. L., Colaizzi, P. D., Evett, S. R., Howell, T. A., and Tolk, J. A.: ET mapping for agricultural water management: present status and challenges, Irrigation Science, 26, 223–236, doi:10.1007/s00271-007-0088-6, 2008.

Gruber, A., Dorigo, W. A., Crow, W., and Wagner, W.: Triple Collocation-Based Merging of Satellite Soil Moisture Retrievals, IEEE Transactions on Geoscience and Remote Sensing, 55, 6780–6792, 2017.

Hagan, M. T. and Menhaj, M.: Training feedforward networks with the Marquardt algorithm, IEEE Trans. Neural Networks, 5, 989–993, 1994.

Hirschi, M., Michel, D., Lehner, I., and Seneviratne, S. I.: A site-level comparison of lysimeter and eddy covariance flux measurements of evapotranspiration, Hydrology and Earth System Sciences, 21, 1809–1825, 2017.

Hobeichi, S., Abramowitz, G., Evans, J., and Ukkola, A.: Derived Optimal Linear Combination Evapotranspiration (DOLCE): a global gridded synthesis ET estimate, Hydrology and Earth System Sciences, 22, 1317–1336, 2018.

Jimenez, C., Prigent, C., Mueller, B., Seneviratne, S. I., McCabe, M. F., Wood, E. F., Rossow, W. B., Balsamo, G., Betts, A. K., Dirmeyer, P. A., Fisher, J. B., Jung, M., Kanamitsu, M., Reichle, R. H., Reichstein, M., Rodell, M., Sheffield, J., Tu, K., and Wang, K.: Global intercomparison of 12 land surface heat flux estimates, Journal of Geophysical Research, 116, D02 102–27, 2011.

Jones, C. S., Finn, J. M., and Hengartner, N.: Regression with strongly correlated data, Journal of Multivariate Analysis, 99, 2136–2153, doi:doi:10.1016/j.jmva.2008.02.008, 2008.

Jung, M., Reichstein, M., Margolis, H. A., Cescatti, A., Richardson, A. D., Arain, M. A., Arneth, A., Bernhofer, C., BONAL, D., Chen, J., Gianelle, D., Gobron, N., Kiely, G., Kutsch, W., Lasslop, G., Law, B. E., Lindroth, A., Merbold, L., Montagnani, L., Moors, E. J., Papale, D., Sottocornola, M., Vaccari, F., and Williams, C.: Global patterns of land-atmosphere fluxes of carbon dioxide, latent heat, and sensible heat derived from eddy

covariance, satellite, and meteorological observations, Journal of Geophysical Research, 116, G00J07–16, 2011.

Kato, T., Tang, Y., Gu, S., Hirota, M., Du, M., Li, Y., and Zhao, X.: Temperature and biomass influences on interannual changes in CO_2 exchange in an alpine meadow on the Qinghai-Tibetan Plateau, Global Change Biology, 12, 1285–1298, doi:10.1111/j.1365-2486.2006.01153.x, 2006.

Kelly, R., Chang, A., Tsang, L., and Foster, J.: A prototype AMSR-E global snow area and snow depth algorithm, IEEE Trans. Geosci. Remote Sens., 41, 230–242, 2003.

Khan, M. S., Liaqat, U. W., Baik, J., and Choi, M.: Stand-alone uncertainty characterization of GLEAM, GLDAS and MOD16 evapotranspiration products using an extended triple collocation approach, Agricultural and Forest Meteorology, 252, 256–268, 2018.

Knohl, A., Schulza, E. D., Kolle, O., and Buchmann, N.: Large carbon uptake by an unmanaged 250-year-old deciduous forest in Central Germany, Agricultural and Forest Meteorology, 118, 151–167, doi:http://dx.doi.org/10.1016/S0168-1923(03)00115-1, 2003.

Le Maitre, D. C. and Versfeld, D. B.: Forest evaporation models: relationships between stand growth and evaporation, Journal of Hydrology, 193, 240–257, 1997.

Lievens, H., Martens, B., Verhoest, N. E. C., Hahn, S., Reichle, R. H., and Miralles, D. G.: Assimilation of global radar backscatter and radiometer brightness temperature observations to improve soil moisture and land evaporation estimates, Remote Sensing of Environment, 189, 194–210, 2017.

Liu, Y. Y., de Jeu, R., and McCabe, M. F.: Global long-term passive microwave satellite-based retrievals of vegetation optical depth, Geophysical . . . , 2011a.

Liu, Y. Y., Parinussa, R. M., Dorigo, W. A., de Jeu, R. A. M., Wagner, W., van Dijk, A. I. J. M., McCabe, M. F., and Evans, J. P.: Developing an improved soil moisture dataset by blending passive and active microwave satellite-based retrievals, Hydrology and Earth System Sciences, 15, 425–436, 2011b.

Ma, S., Baldocchi, D. D., Xu, L., and Hehn, T.: Inter-annual variability in carbon dioxide exchange of an oak/grass savanna and open grassland in California, Agricultural and Forest Meteorology, 147, 157–171, doi:10.1016/j.agrformet.2007.07.008, 2007.

Martens, B., Miralles, D. G., Lievens, H., van der Schalie, R., de Jeu, R. A. M., Férnandez-Prieto, D., Beck, H. E., Dorigo, W. A., and Verhoest, N. E. C.: GLEAM v3: satellite-based land evaporation and root-zone soil moisture, Geoscientific Model Development Discussions, 0, 1–36, 2016.

McCabe, M. F., Ershadi, A., Jimenez, C., Miralles, D. G., Michel, D., and Wood, E. F.: The GEWEX LandFlux project: evaluation of model evaporation using tower-based and globally gridded forcing data, Geoscientific Model Development, 9, 283–305, 2016.

McCaughey, J. H., Pejam, M. R., Arain, M. A., and Cameron, D. A.: Carbon dioxide and energy fluxes from a boreal mixedwood forest ecosystem in Ontario, Canada, Agricultural and Forest Meteorology, 140, 79–96, doi:10.1016/j.agrformet.2006.08.010, 2006.

McEwing, K. R., Fisher, J. P., and Zona, D.: Environmental and vegetation controls on the spatial variability of CH4 emission from wet-sedge and tussock tundra ecosystems in the Arctic, Plant and soil, 388, 37–52, 2015.

Michel, D., Jimenez, C., Miralles, D. G., Jung, M., Hirschi, M., Ershadi, A., Martens, B., McCabe, M. F., Fisher, J. B., MU, Q., Seneviratne, S. I., Wood, E. F., and Fernández-Prieto, D.: The WACMOS-ET project –

Part 1: Tower-scale evaluation of four remote-sensing-based evapotranspiration algorithms, Hydrology and Earth System Sciences, 20, 803–822, 2016.

Milyukova, I. M., Kolle, O., Varlagin, A. V., Vygodskaya, N. N., Schulze, E. D., and Lloyd, J.: Carbon balance of a southern taiga spruce stand in European Russia, Tellus B, 54, 429–442, doi:10.1034/j.1600-0889.2002.01387.x, 2002.

Miralles, D. G., Holmes, T. R. H., de Jeu, R. A. M., Gash, J. H., Meesters, A. G. C. A., and Dolman, A. J.: Global land-surface evaporation estimated from satellite-based observations, Hydrology and Earth System Sciences, 15, 453–469, 2011.

Miralles, D. G., Jimenez, C., Jung, M., Michel, D., Ershadi, A., McCabe, M. F., Hirschi, M., Martens, B., Dolman, A. J., Fisher, J. B., MU, Q., Seneviratne, S. I., Wood, E. F., and Fernández-Prieto, D.: The WACMOS-ET project - Part 2: Evaluation of global terrestrial evaporation data sets, Hydrology and Earth System Sciences, 20, 823–842, 2016.

Moncrieff, J., Malhi, Y., and Leuning, R.: The propagation of errors in long-term measurements of land-atmosphere fluxes of carbon and water, Global Change Biology, 2, 231–240, doi:10.1111/j.1365-2486.1996.tb00075.x, http://dx.doi.org/10.1111/j.1365-2486.1996.tb00075.x, 1996.

Monteith, J.: Evaporation and environment, Symp. Soc. Exp. Biol, 19, 4, 1965.

Moureaux, C., Debacq, A., Bodson, B., Heinesch, B., and Aubinet, M.: Annual net ecosystem carbon exchange by a sugar beet crop, Agricultural and Forest Meteorology, 139, 25–39, 2006.

Mu, Q., Zhao, M., and Running, S. W.: Improvements to a MODIS global terrestrial evapotranspiration algorithm, Remote Sensing of Environment, 115, 1781–1800, 2011.

Mueller, B., Seneviratne, S. I., Jimenez, C., Corti, T., Hirschi, M., Balsamo, G., Ciais, P., Dirmeyer, P., Fisher, J. B., Guo, Z., Jung, M., Maignan, F., McCabe, M. F., Reichle, R., Reichstein, M., Rodell, M., Sheffield, J., Teuling, A. J., Wang, K., Wood, E. F., and Zhang, Y.: Evaluation of global observations-based evapotranspiration datasets and IPCC AR4 simulations, Geophysical Research Letters, 38, n/a–n/a, 2011.

Mueller, B., Hirschi, M., Jimenez, C., Ciais, P., Dirmeyer, P. A., Dolman, A. J., Fisher, J. B., Jung, M., Ludwig, F., Maignan, F., Miralles, D. G., McCabe, M. F., Reichstein, M., Sheffield, J., Wang, K., Wood, E. F., Zhang, Y., and Seneviratne, S. I.: Benchmark products for land evapotranspiration: LandFlux-EVAL multi-data set synthesis, Hydrology and Earth System Sciences, 17, 3707–3720, 2013.

Munier, S. F. A. S. S. C. P. F. P. P. M. and Pan, M.: Combining data sets of satellite-retrieved products for basin-scale water balance study: 2. Evaluation on the Mississippi Basin and closure correction model, pp. 1–17, 2014.

Nguyen, D. and Widrow, B.: Improving the learning speed of 2-layer neural networks by choosing initial values of the adaptative weights, in: Proceedings of the 1990 International Joint Conference on Neural Networks, pp. 21–26, 1990.

Noormets, A., McNulty, S. G., DeForest, J. L., Sun, G., Li, Q., and Chen, J.: Drought during canopy development has lasting effect on annual carbon balance in a deciduous temperate forest, New Phytologist, 179, 818–828, doi:10.1111/j.1469-8137.2008.02501.x, 2008.

Nordbo, A., Järvi, L., and Vesala, T.: Revised eddy covariance flux calculation methodologies – effect on urban energy balance, Tellus B: Chemical and Physical Meteorology, 64, 18 184, 2012.

Pauwels, V. R. N., Timmermans, W., and Loew, A.: Comparison of the estimated water and energy budgets of a large winter wheat field during AgriSAR 2006 by multiple sensors and models, Journal of Hydrology, 349, 425–440, 2008.

Pinty, B., Lavergne, T., Vossbeck, M., Kaminski, T., Aussedat, O., Giering, R., Gobron, N., Taberner, M., Verstraete, M. M., and Widlowski, J.-L.: Retrieving surface parameters for climate models from Moderate Resolution Imaging Spectroradiometer (MODIS)-Multiangle Imaging Spectroradiometer (MISR) albedo products, Journal of Geophysical Research: Atmospheres, 112, doi:10.1029/2006JD008105, 2007.

Pinty, B., Jung, M., Kaminski, T., Lavergne, T., Mund, M., Plummer, S., Thomas, E., and Widlowski, J.: Evaluation of the JRC-TIP 0.01° products over a mid-latitude deciduous forest site, Remote Sens. Environ., 115, 3567–3581, 2011a.

Pinty, B., Taberner, M., Haemmerle, V., Paradise, S., Vermote, E., Verstraete, M., Gobron, N., and Widlowski, J.-L.: Global-Scale Comparison of MISR and MODIS Land Surface Albedos, J Climate, 24, 732–749, 2011b.

Post, H., Hendricks Franssen, H. J., Graf, A., Schmidt, M., and Vereecken, H.: Uncertainty analysis of eddy covariance CO₂ flux measurements for different EC tower distances using an extended two-tower approach, Biogeosciences, 12, 1205–1221, 2015.

Priestley, C. and Taylor, R.: On the assessment of surface heat flux and evaporation using large-scale parameters, Mon. Weather Rev, Mon. Weather Rev., 100, 81–92, 1972.

Rambal, S., Joffre, R., Ourcival, J. M., Cavender-Bares, J., and Rocheteau, A.: The growth respiration component in eddy CO2 flux from a Quercus ilex mediterranean forest, Global Change Biology, 10, 1460–1469, 2004.

Richardson, A. D., Hollinger, D. Y., Burba, G. G., Davis, K. J., Flanagan, L. B., Katul, G. G., Munger, J. W., Ricciuto, D. M., Stoy, P. C., and Suyker, A. E.: A multi-site analysis of random error in tower-based measurements of carbon and energy fluxes, Agricultural and Forest Meteorology, 136, 1–18, 2006.

Rodgers, C. D.: Inverse methods for atmospheric sounding: Theory and practise. Series on Atmospheric, Oceanic and Planetary Physics, vol. 2, World Scientific Publishing, 1 edn., 2000.

Schmid, H. P., Grimmond, C. S. B., Cropley, F., Offerle, B., and Su, H. B.: Measurements of CO2 and energy fluxes over a mixed hardwood forest in the mid-western United States, Agricultural and Forest Meteorology, 103, 357–374, doi:10.1016/S0168-1923(00)00140-4, 2000.

Scott, R. L., Jenerette, G. D., Potts, D. L., and Huxman, T. E.: Effects of seasonal drought on net carbon dioxide exchange from a woody-plant-encroached semiarid grassland, Journal of Geophysical Research: Biogeosciences, 114, G04 004, 2009.

Scott, R. L., Jenerette, G. D., Potts, D. L., and Huxman, T. E.: Effects of seasonal drought on net carbon dioxide exchange from a woody-plant-encroached semiarid grassland, Journal of Geophysical Research-Biogeosciences, 114, doi:10.1029/2008JG000900, 2009.

Simbahan, G. C., Dobermann, A., Goovaerts, P., Ping, J. L., and Haddix, M. L.: Fine-resolution mapping of soil organic carbon based on multivariate secondary data, Geoderma, 132, 471–489, doi:10.1016/j.geoderma.2005.07.001, 2006.

Sorooshian, S., Lawford, R., and Try, P.: Water and energy cycles: Investigating the links, WMO Bulletin, 54, 58–64, 2005.

Stackhouse, P., Gupta, S., Cox, S., Mikovitz, J., Zhang, T., and Chiacchio, M.: 12-year surface radiation budget data set, GEWEX News, 14, 10–12, 2004.

Steininger, M. K.: Net carbon fluxes from forest clearance and regrowth in the Amazon, Ecological Applications, 14, 313–322, doi:10.1890/02-6007, 2004.

Takala, M., Luojus, K., Pulliainen, J., Derksen, C., Lemmetyinen, J., Kärnä, J.-P., Koskinen, J., and Bojkov, B.: Estimating northern hemisphere snow water equivalent for climate research through assimilation of space-borne radiometer data and ground-based measurements, Remote Sensing of Environment, 115, 3517–3529, 2011.

Talsma, C. J., Good, S. P., Jimenez, C., Martens, B., Fisher, J. B., Miralles, D. G., McCabe, M. F., and Purdy, A. J.: Partitioning of evapotranspiration in remote sensing-based models, Agricultural and Forest Meteorology, 260-261, 131–143, 2018.

Twine, T. E., Kustas, W. P., Norman, J. M., Forest, D. C. A., , and 2000: Correcting eddy-covariance flux underestimates over a grassland, Elsevier, 103, 279–300, 2000.

Verma, S. B., Dobermann, A., Forest, K. C. A., , and 2005: Annual carbon dioxide exchange in irrigated and rainfed maize-based agroecosystems, Elsevier, 131, 77–96, 2005.

Verma, S. B., Dobermann, A., Cassman, K. G., Walters, D. T., Knops, J. M., Arkebauer, T. J., Suyker, A. E., Burba, G. G., Amos, B., Yang, H. S., Ginting, D., Hubbard, K. G., Gitelson, A. A., and Walter-Shea, E. A.: Annual carbon dioxide exchange in irrigated and rainfed maize-based agroecosystems, Agricultural and Forest Meteorology, 131, 77–96, doi:10.1016/j.agrformet.2005.05.003, 2005.

Vinukollu, R. K., Meynadier, R., Sheffield, J., and Wood, E. F.: Multi-model, multi-sensor estimates of global evapotranspiration: climatology, uncertainties and trends, Hydrological Processes, 25, 3993–4010, 2011.

Wang, J., Zhuang, J., Wang, W., Liu, S., and Xu, Z.: Assessment of Uncertainties in Eddy Covariance Flux Measurement Based on Intensive Flux Matrix of HiWATER-MUSOEXE, IEEE Geoscience and Remote Sensing Letters, 12, 259–263, 2015.

Wang, K. and Dickinson, R. E.: A review of global terrestrial evapotranspiration: Observation, modeling, climatology, and climatic variability, Reviews of Geophysics, 50, RG2005–54, 2012.

Willmott, C. J.: Some comments on the evaluation of model performance, Bulletin of the American Meteorological Society, 63, 1309–1313, 1982.

Wilson, K., Goldstein, A., Falge, E., Forest, M. A. A., , and 2002: Energy balance closure at FLUXNET sites, Elsevier, 113, 223–243, 2002.

Yao, Y., Liang, S., Li, X., Chen, J., Liu, S., Jia, K., Zhang, X., Xiao, Z., Fisher, J. B., Mu, Q., Pan, M., Liu, M., Cheng, J., Jiang, B., Xie, X., Gr nwald, T., Bernhofer, C., and ROUPSARD, O.: Improving global terrestrial evapotranspiration estimation using support vector machine by integrating three process-based algorithms, Agricultural and Forest Meteorology, 242, 55–74, 2017.

Yilmaz, M. T., Crow, W. T., Anderson, M. C., and Hain, C.: An objective methodology for merging satellite- and model-based soil moisture products, Water Resources Research, 48, W11 502, 2012.

Zeeman, M. J., Hiller, R., Gilgen, A. K., Michna, P., Plüss, P., Buchmann, N., and Eugster, W.: Management, not climate, controls net CO_2 fluxes and carbon budgets of three grasslands along an elevational gradient in Switzerland, Agricultural Forest Meteorology, 50, 519–530, 2010.

Zhang, K., Kimball, J. S., and Running, S. W.: A review of remote sensing based actual evapotranspiration estimation, Wiley Interdisciplinary Reviews: Water, 3, 834–853, 2016.