Rebuttal BG-2021-355

Local scale evaluation of the simulated interactions between energy, water and vegetation in land surface model

Jan De Pue, José Miguel Barrios, Liyang Liu, Philippe Ciais, Alirio Arboleda, Rafiq Hamdi, Manuela Balzarolo, Fabienne Maignan, and Françoise Gellens-Meulenberghs

Handling Associate Editor: Ivonne Trebs

May 19, 2022

Anonymous referee 2

Three models (prognostic models ISBA and ORCHIDEE and a diagnostic model rooted in satellite remote sensing product development) were run on the meteorological and land surface forcing data of 56 eddy-covariance sites to compare their results to each other and to the site observations. Results were compared with multiple strategies (including bias vs. RMSE, Taylor diagrams, sensitivities between state variables per land use type, error correlations, phenology and seasonal cycles) and results were used to identify (where possible) or hyothesize (otherwise) weaknesses of the different models, also considering uncertainties in the observation data. Main findings include a better performance of the diagnostic as compared to teh prognostic models, a convergence in strengths and weaknesses between both prognostic models partly due to latest updates of the ISBA model, but also remaining differences, and recommendations on most important future improvements (in particular related to drought stress response, phenology and biomass allocation). The manuscript is very well written, methods and results are presented in clarity, the subject is relevant to Biogeosciences, and original among others in the sense that the newest version of ISBA and a comparatively new data product of teh eddy-covariance network are used.
Notwithstanding a lack of expertise on my side when it comes to internal details of the used models, which would ideally be addressed by other reviewers, I recommend the manuscript for publication after minor revisions suggested below.
Detailed Comments:

Comment 3.1 — Title: Consider adding "three" before land surface models, currently is lets readers easily think of a large multi-model study.

> Agreed, the title was changed accordingly

Comment 3.2 — Figure 1: Relation and feedbacks in the caption is no distinction that makes it easy to understand for the reader - relation could also be something purely empirical but here apparently you mean it as "more direct / stronger / first order" than the feedbacks. Check if this distinction is really needed and if, which other words

could stress it. Next line of the caption, from the arrow it would be better to write Soil moisture - LAI than vice versa. Figure itself: Would it make sense to add stomatal control somewhere in the middle? The way it is now it seems like LE and GPP are each controlled independently by soil moisture and LAI, such that the reader is almost wondering why there is not also an arrow between them.

The caption was corrected following the recommendations of the reviewer:

> *" First order relations (plain lines) and feedbacks (dashed lines) of the state variables and surface fluxes in prognostic LSM. The feedback mechanisms are not present in diagnostic models, and the Soil moisture-LAI relation (dotted line) occurs only in prognostic LSM with dedicated phenology schemes. "*

Adding the stomatal control in the middle would make sense for the prognostic models. The soil moisture modulates the stomatal closure, which in turn affects LE and GPP. However, we choose not to display it as such, as this would not be an accurate representation of the mechanisms in the diagnostic model. Instead, we prefer to keep the more general schematic.

Comment 3.3 — L97: "corrected manually to represent the tower footprint area" is somewhat unclear, could you be more specific?

This was rephrased:

> *" The land cover at each site was derived from ECOCLIMAP 2 (Faroux et al., 2013) and corrected manually if this was not representative for the tower footprint area (based on ICOS and FLUXNET metadata, and satellite imagery). "*

Comment 3.4 — L99: "linearly interpolated" refers to the ERA5 data being hourly and the tower observations mostly half-hourly?

Indeed, the text was revised:

> *" The forcing from ERA5 (hourly resolution) was linearly interpolated to match the 30 minute temporal resolution from the tower observations. "*

Comment 3.5 — L140 & 159: Could the free drainage at the bottom explain the over-sensitivity to drought stress? (To be discussed not here)

Indeed, the lack of ground water dynamics might impact the results. This was referred to in the discussion section:

> *" The local scale simulations in this study were not coupled to a hydrological model, thus ground water dynamics were lacking. Though only a limited effect of capillary rise was found in studies with a coupled groundwater hydrology, the impact can be non-negligible for forest ecosystems with a deep root system (Decharme et al., 2019; MacBean et al., 2020). The further development of ground water dynamics in LSM is indispensable for the accurate coupling of energy, water and carbon in forest vegetation and its response to severe drought events. "*

The free drainage boundary condition might result in an overestimation of the occurrence of water-limited conditions.

However, in the sensitivity analysis we only consider the relation between anomalies in the simulated soil moisture and the fluxes. In principle, the sensitivity of this analysis to the frequency of water-limited conditions should be limited.
If there would be an issue caused by the free drainage boundary condition, it would be primarily visible in the overestimated frequency of water-limited conditions.

Comment 3.6 — L161 & 203: Just mentioning "a selection of sites [...] to ensure adequate data quality" is a bit arbitrary

To clarify the site selection process, the methods & material section was revised:

> *" From the FLUXNET2015 dataset (Pastorello et al., 2020) and the ICOS '2018 drought initiative' dataset (Drought 2018 Team and ICOS Ecosystem Thematic Centre, 2019), sites were selected with adequate data quality (at least 1 year of carbon fluxes, dominated by observations with quality flag 1 or better), homogeneous land cover and limited disturbance due to management. This resulted in the 56 sites, listed in Table 2, and a total of 526 simulation years. 33 of these sites are dominated by forest land cover, whereas 18 are dominated by herbaceous vegetation and 5 are crop sites (the models are configured to run without management practices). The FLUXNET and ICOS data products had been pre-processed with the ONEFLUX processing pipeline (Pastorello et al., 2020). "*

Comment 3.7 — L165: Again of course not to be discussed here, could the non-specified management practices be an important explanation for the difference between diagnostic and prognostic model(s) in crop sites (e.g. Figure 3)?

Yes, some of the management practices are implicitly present in the forcing of the diagnostic model, whereas they are missing in the prognostic models.
However, the performance of the prognostic models in crop sites is not significantly different to that in natural herbaceous sites. From these results, it is not clear to what extent missing management practices have an impact on the results.
We add the following to the discussion:

> *" In the crop sites, management practices were missing in the prognostic models. In the mean annual cycle of LAI (Fig. 11), it is evident that the no harvest occurs. Despite this, the simulations of LE were not significantly less accurate compared to other land cover types. After harvest, LE consists largely out of bare soil evaporation. Though vegetation was still present in the models, the bulk LE was still reasonably accurate. More evident degradation of the results was found in GPP after harvest, which was overestimated. Even in the diagnostic model, where management practices were incorporated implicitly in the forcing variables, GPP was overestimated. Notably, despite the missing management practices in the prognostic models, the quality of the simulated LE and GPP (and their anomalies) was not significantly different from that in natural herbaceous sites. "*

Comment 3.8 — L168: Were the PFT and vegetation type info derived from the IGBP metadata of the flux network, or from remote sensing, or other sources?

> The PFT was derived from the IGBP metadata of FLUXNET/ICOS. The text was revised:
>
> > *" The test sites were classified per PFT (taken from the FLUXNET/ICOS IGBP metadata),... "*

Comment 3.9 — Table 2: Some sites (apparently especially crop sites, e.g. BE-Lon, DE-Kli and DE-RuS) are listed with very large LE corrections, that do not match what I thought I knew from past studies on their energy balance closure. I tried a detailed check on DE-RuS: The flux-weighted effective average factor between LE_corr and LE is 1.44, somewhat lower (why) than the 1.47 in Table 2, but still far too high compared to any energy balance analysis carried out for this site in the past (e.g. 1.18 would result from Eder et al. 2015, DOI: 10.1175/JAMC-D-14-0140.1 which focused on summer months and 1.23 from Graf et al. 2020, doi.org/10.1098/rstb.2019.0524 with a study period matching the drought2018 dataset). Note that the current One-Flux product does to my knowledge not use details such as heat flux plate depth important to compute the energy balance closure, however even considering this the difference seems far too large, so it seems that LE_corr from this dataset should be used with care.

> The value of LE_corr was calculated using aggregated daily LE values (LE_CORR and LE_F_MDS). The reported value in the table is the mean (0.470523). If the median is used, we find 0.44 (0.437974). Perhaps that is the difference.
> The calculation considers the full time period, and doesn't exclude the winter period, (during which LE is smaller and the relative correction is higher, see also Fig. S-1).
> The uncertainty associated with the eddy covariance observations is discussed elaborately throughout this manuscript (e.g. section 4.1). We fully agree that the validation results need to be interpreted with care.

Comment 3.10 — L182: Actually ICOS does have a standardized setup for soil moisture; however, the used dataset (drought2018) still mostly consists of so-called "legacy data" (i.e. voluntarily provided measurements with pre-ICOS set-ups). No need to mention it, just avoid the misleading wording.

> We were not aware of this, good to know for the future!
> The sentence was changed as follows:
>
> > *" Not all sites are equipped with soil moisture sensors, nor is there a standardized setup or post-processing for soil moisture in the datasets used for this study. "*

Comment 3.11 — L196: From the way it is mentioned for LE and H and then a new paragraph starts, no EBC-based correction was assumed for NEE (and propagated to GPP and Reco)? Not that I would like to recommend it, just for clarity. Unfortunately even the correctability for LE and H is far from certain, but then depending on the assumed reasons it may or may not also apply to the CO2 flux (at least its turbulent part before WPL correction). It is nothing that can be done in a more certain way, but it is important to be aware of it later e.g. when LE and NEE show different model-observation biases. P.S.: It is nicely mentioned already in line 394, but may still leave the reader wondering here.

> Indeed. For clarity we mention it here as well:

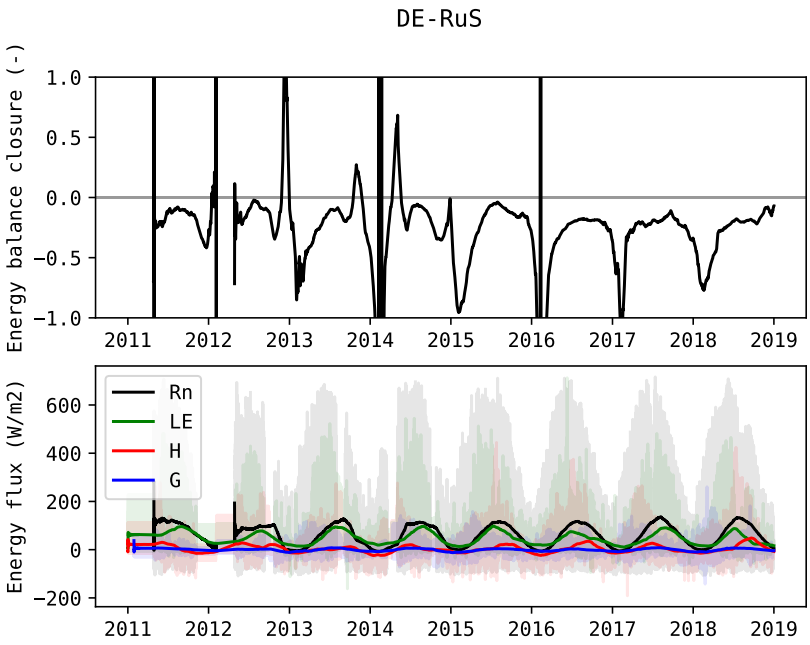

Figure S-1: Energy balance at DE-RuS

> *" Though some authors have recommended to correct the carbon fluxes in a similar way as the turbulent fluxes, such a procedure was not included in the processing pipeline (Massman and Lee, 2002; Gao et al., 2019, see also section 4). "*

Comment 3.12 — L221: Make clearer if the mean annual cycles are computed one per site across all its site-years (which implies that the deviations also include inter-annual variability, which is not a bad thing but one to be aware of)

The text was revised to clarify this, following the recommendation of the reviewer:

> *" The mean annual cycles were computed per site, across all its site-years. "*

Comment 3.13 — L226-235 and Figures 5+6: Better explain for what the slope and for what the correlation was used. Comparing the text to the figure captions, I guess that the "Spearman slope" in the caption is wrong (slope yes but probably not between the rank-transformed variables, which is what "Spearman" would imply to me)?

There was indeed a mistake in the captions of the figures 5 and 6. "Spearman slope" should have been "slope". This was corrected.
The slope was used to evaluate the response of the surface fluxes to soil moisture and LAI. The analysis was done for the observations and the models. This allows to evaluate how well the sensitivity in the models resembles the observed sensitivity. It is explained in the manuscript as follows:

> *" To assess the sensitivity of the fluxes to the state variables (S_e and LAI), the slope of the seasonal anomalies of the fluxes against the anomalies of the state variables was determined. This analysis was performed for the observations and the simulations, and compared. Note that the linear slope was used here, though a linear response is not necessarily expected (e.g. the response to soil moisture anomalies depend on a wet/dry regime). The goal of this analysis was to investigate whether LSM are capable of reproducing a similar relationship as found in the observations. Significant differences between the models were evaluated with the Wilcoxon signed-rank test. "*

Additional explanation was given in the result section, to help the reader:

> *" The sensitivity of the surface fluxes to soil moisture and LAI was quantified with a simple linear regression between their anomalies. The slope of these regressions indicates the strength of the response to the state variables. "*

The error correlation was used to evaluate whether errors in the surface fluxes are associated with errors in the state variables (soil moisture and LAI).

> *" To evaluate whether errors in the state variables are associated with errors in the surface fluxes (or vice versa), the Spearman rank correlation between both was calculated. "*

Similarly, this was mentioned again in the result section:

> *" To evaluate the impact of the quality of S_e and LAI on the simulated surface fluxes, the Spearman correlation of the errors in the state variables and the fluxes was calculated. "*

Comment 3.14 — L243: Maybe adding "independently" to the last sentence and mentioning it already at the start of subsection 2.3.2 would make it easier to understand.

This section was restructured to improve readability.

Comment 3.15 — L247: Here it is unclear whether the LE partitioning methods are just mentioned out of interest, or were applied in this study (which seems not to be the case according to the result section).

The transpiration derived from the observations was used to estimate the WUE in the sites. This was clarified:

> *" From the GPP and transpiration (Tr), the WUE was derived:*
>
> $$WUE = \frac{GPP}{Tr} \tag{S-1}$$
>
> *"*

Comment 3.16 — L255: Shouldn't this be visible in Fig. 2a? If accuracy corresponds with the bias (x axis) and precision with random errors (y axis), it would be more accurate to state that both models have the same accuracy but ISBA a slightly better precision. Sometimes accuracy is also used as a combined name corresponding to both,

systematic and random errors; then the statement is true but imprecise and the "significantly" seems a bit overstated (unless it refers to a successful statistical significance test of course).

Some confusion might arise from the use of the words "bias" and "accuracy". The terminology as it is used in this manuscript was defined at L250:

" The bias (ME) and accuracy (RMSE) of the simulated... "

To our knowledge it is not unusual to use bias and accuracy in this sense. Significantly refers indeed to a statistical significant difference with the Wilcoxon test. Note that this test is a pairwise comparison. These significant differences might not be evident from a figure like Fig. 2a. A scatter plot reveals this difference more clearly (see Fig. S-2).

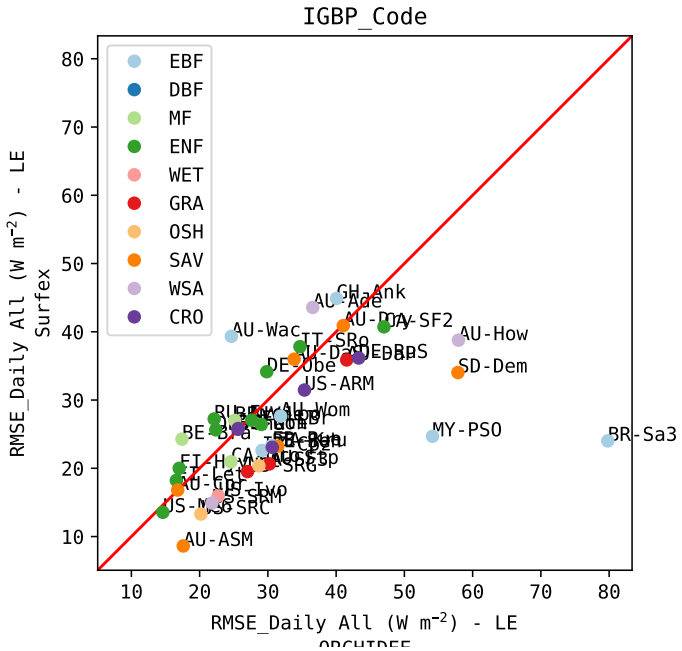

Figure S-2: RMSE of the simulated LE with ISBA and ORCHIDEE.

Comment 3.17 — Around L350, Figure 12: Are differences in the partitioning between drainage and runoff really interesting to discuss for models which were all run in uncoupled 1D mode? My (maybe wrong) expectation would be that it is a quite arbitrary function of model physics that only converges between models if horizontal neighbours with given slopes get a chance to communicate with each other.

The primary goal of showing the water balance is to provide further insight in the simulated water dynamics. Within the frame of this 1D experiment, the difference in water partitioning are deemed relevant, even though the models might behave differently in a coupled mode.
It is not clear why we could assume that the models would converge in a coupled/3D experiment.

Comment 3.18 — L389: "caused by surface" looks a bit as if something was missed out here, maybe "... surface heterogeneities"?

Indeed, this was corrected.

Comment 3.19 — L396: Mentioning GPP and LE alongside each other with parenthesis does not fully capture the extend of the problem (see also comment on L196): Since LE was corrected for EBC non-closure (at least tried to, given the open questions correctly mentioned by the authors) while GPP was not, it could somewhat be expected that the mean difference model vs. obs is smaller (or more negative) for LE and larger (or zero or less negative) for GPP. This is exactly what we see in Fig. 2 (if the x axis is model - observation). Which might indicate (among other ways to explain more associated with model shortcomings of course) that the LE is overcorrected even by the current EBC correction. Note that to my knowledge (if I didn't overlook something) Gebler et al. (2015) do not report a better EC-lysimeter match by putting the whole deficit into LE, but by a correction conserving the evaporative fraction, which is similar to the Bowen ratio conserving correction by Pastorello et al. Others even suggest that most or all of the deficit might be related to sensible heat (Ingwersen et al. 2011, https://doi.org/10.1016/j.agrformet.2010.11.010), or found a good match with independent reference data without LE correction (e.g. Graf et al. 2014, https://doi.org/10.1002/2013WR0 for the catchment water budget method). In general, a problem with the body of existing comparisons of eddy-covariance fluxes to independent reference methods is that the latter can have their own systematic errors (e.g. island effect in case of lysimeters, or different footprints of both systems) on a similar order of magnitude as the eddy-covariance energy balance closure gap, and that the (often quite definite) answers of the single studies are in conflict when comparing these studies with each other. Maybe (especially given the risk of a too large energy balance gap seen by the flux product as discussed in comment on table 2) it would even be interesting to see how the model-observation match without the energy balance correction is. Of course, the results would not completely reliably indicate an overcorrection / different source of the closure gap, but could also point to an unintended adaption of the models towards uncorrected eddy-covariance data during past validations.

We fully agree with this comment, highlighting the uncertainty related to the eddy covariance observations. The suggested additional references and discussion was incorporated in the text.

> *" Furthermore, some studies have indicated that the eddy covariance observations are closer to lysimeter data if the energy balance is closed by correcting LE only (Wohlfahrt et al., 2010). Considering this, the negative bias of the simulated LE (and GPP) in this study could be even underestimated. Conversely, others suggest that most or all of the deficit might be related to H (Ingwersen et al., 2011), or found a good match with independent reference data without LE correction (Graf et al., 2014). Validation results of the turbulent fluxes without energy balance closure correction are given in the supplement material. "*

In addition, we provide some plots of the validation with and without EBC correction in the supplementary material (see Fig. S-3). Finally, the reference to Gebler et al. (2015) was indeed false. It should have been a reference to Wohlfahrt et al. (2010). This was corrected.

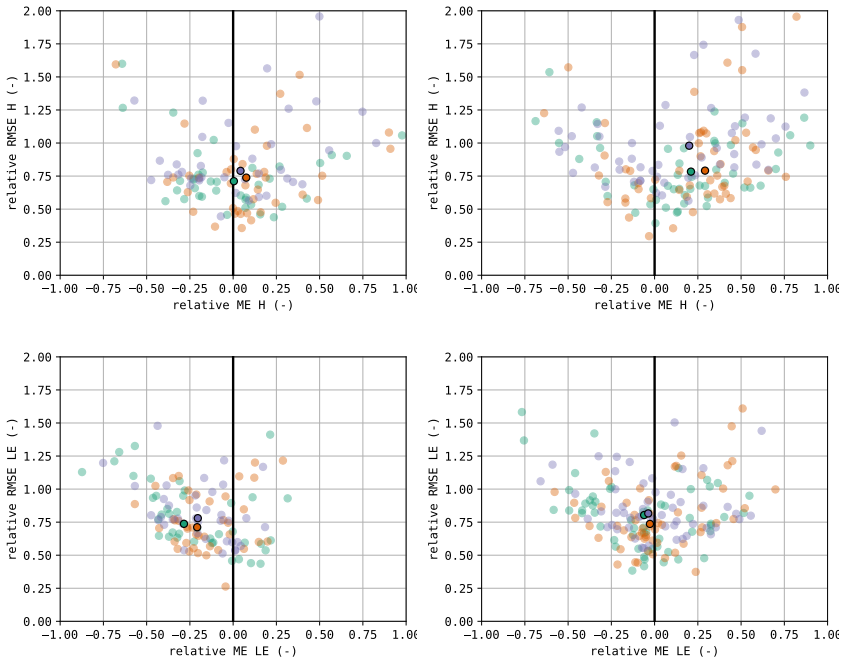

Figure S-3: Validation of H and LE (top and bottom row, respectively), with and without EBC correction (left and right column, respectively).

Comment 3.20 — L399-407 (4.1.2): Almost all differences discussed here could also be due to the different management intensity between forests and herbaceous, the latter including the crop sites (I guess?) and also intensively managed grassland. The prognostic models were not informed about management (e.g. which crop, when bare soil), while for the diagnostic model some of teh effects of management may have been implicit in the data provided.

We have modified the text to clarify that "herbaceous sites" refers to sites, dominated by short vegetation and limited management. They do not include crop sites.

> *" Generally, the differences between the accuracy of the simulated surface fluxes was most distinct in the sites dominated by herbaceous vegetation (excluding crop sites). "*

Additionally, some extra discussion on the results related to the crop sites and management was added (see point 7).

Comment 3.21 — L436: Do not understand why this (slow buildup of LAI in early season) should be a consequence of the sentence before (assimilated carbon invested into leaves first). Do you maybe mean the same thing you describe more understandably in the next paragraph, i.e. that a process is missing in ISBA which can grow leaves from stored biomass?

Yes, this was rephrased as follows:

> *" The assimilated carbon is attributed to the leaf biomass pool first, from where it trickles down to the other pools. No carbon reserve*

> *dynamics are implemented. The consequence is that the simulated LAI in ISBA starts slow during spring, as GPP is underestimated due to a low LAI. It continues to build up LAI until late in the second half of the season, when photosynthetic conditions become sub-optimal, and leaf senescence is triggered. In contrast, the observed seasonal LAI cycles reach a maximum in the first half of the growing season. "*

Comment 3.22 — L516-519: Maybe for readers jumping to the conclusions section it would be helpful to give a brief hint on the most important process(es) underrepresented, e.g. as discussed around L446-452 (where does the biomass for new leaves come from at season start).

> The following sentence was added:
>
> *" Processes describing carbon reserve dynamics during spring and leaf senescence were found to be falling short or missing. "*

Comment 3.23 — Acknowledgement: despite much praise ("This work stands on the shoulders. . . ") and correctly citing the DOI of the drought2018 data product, the attribution of the work at least of the flux site PIs offering the flux data is a bit awkward. The explanation attached to data policy states that PIs should be contacted before publication (to learn about possible acknowledgement requirements, or in extreme cases offer the possibility to scientifically contribute to the study) at least in case the data play a very substantial role for the publication. It might be argued that the latter is the case here. I am well aware that the current situation is unsatisfying for both sides (study authors cannot continue forever to ask hundreds of data authors for each multi-site synthesis, which would delay scientific progress and encourage bagatelle coauthorships; but the latter still often feel incompletely compensated for their voluntary work, given that in most countries they are unfortunately not paid for the site servicing and raw data processing the way weather service employees are, but for science, often on non-permanent contracts, from which they divert worktime for the data production), and do not suggest to revise the communication workflow for this study, but would like to remind the authors and community of it for future studies - at least until either DOI citations have become a highly valued measure of recognition, or data providers are mainly employed to provide free data and most of the data processing including raw data to flux processing has been taken over by the central facilities of the next-generation networks.

> We fully agree with this remark. PIs were indeed not contacted prior to publication, though their datasets are indispensable for this study. It is too late to change this now, but we will keep this in mind for future work.

Comment 3.24 — Purely Technical Comments:
L23: not sure "to better" is good English, maybe "improve"?
L65 & 105: Check if "remote sensed" works, maybe "remotely sensed" or "remote-sensing based" (could also be satellite based if it is exclusively satellites)
L307: "In" missing before "Fig. 6"
L330: error*s*?
L362: frequent*ly*
L367: check usage of "in/on(?) the one / other hand"

L467: Blank missing at start of new sentence.
L470: Maybe replace "Which" by "This"
L471: "wears many hats", just like the above, maybe a bit too "oral" style.
L479: "shows" instead of "learns"?

> Most of these recommended changes were adopted.

References

B. Decharme, C. Delire, M. Minvielle, J. Colin, J.-P. Vergnes, A. Alias, D. Saint-Martin, R. Séférian, S. Sénési, and A. Voldoire. Recent changes in the isba-ctrip land surface system for use in the cnrm-cm6 climate model and in global off-line hydrological applications. *Journal of Advances in Modeling Earth Systems*, 11(5):1207–1252, 2019.

Drought 2018 Team and ICOS Ecosystem Thematic Centre. Drought-2018 ecosystem eddy covariance flux product in fluxnet-archive format - release 2019-1, 2019. URL `https://meta.icos-cp.eu/collections/UZw8ra7OVilmVjATTCgIimpz`.

S. Faroux, A. Kaptué Tchuenté, J.-L. Roujean, V. Masson, E. Martin, and P. L. Moigne. Ecoclimap-ii/europe: A twofold database of ecosystems and surface parameters at 1 km resolution based on satellite information for use in land surface, meteorological and climate models. *Geoscientific Model Development*, 6(2):563–582, 2013.

Z. Gao, H. Liu, J. E. Missik, J. Yao, M. Huang, X. Chen, E. Arntzen, and D. P. Mcfarland. Mechanistic links between underestimated co2 fluxes and non-closure of the surface energy balance in a semi-arid sagebrush ecosystem. *Environmental Research Letters*, 14(4):044016, 2019.

S. Gebler, H.-J. Hendricks Franssen, T. Pütz, H. Post, M. Schmidt, and H. Vereecken. Actual evapotranspiration and precipitation measured by lysimeters: a comparison with eddy covariance and tipping bucket. *Hydrology and earth system sciences*, 19(5):2145–2161, 2015.

A. Graf, H. R. Bogena, C. Drüe, H. Hardelauf, T. Pütz, G. Heinemann, and H. Vereecken. Spatiotemporal relations between water budget components and soil water content in a forested tributary catchment. *Water resources research*, 50(6):4837–4857, 2014.

J. Ingwersen, K. Steffens, P. Högy, K. Warrach-Sagi, D. Zhunusbayeva, M. Poltoradnev, R. Gäbler, H.-D. Wizemann, A. Fangmeier, V. Wulfmeyer, et al. Comparison of noah simulations with eddy covariance and soil water measurements at a winter wheat stand. *Agricultural and Forest Meteorology*, 151(3):345–355, 2011.

N. MacBean, R. L. Scott, J. A. Biederman, C. Ottlé, N. Vuichard, A. Ducharne, T. Kolb, S. Dore, M. Litvak, and D. J. Moore. Testing water fluxes and storage from two hydrology configurations within the orchidee land surface model across us semi-arid sites. *Hydrology and Earth System Sciences*, 24(11):5203–5230, 2020.

W. Massman and X. Lee. Eddy covariance flux corrections and uncertainties in long-term studies of carbon and energy exchanges. *Agricultural and Forest Meteorology*, 113(1-4):121–144, 2002.

G. Pastorello, C. Trotta, E. Canfora, H. Chu, D. Christianson, Y.-W. Cheah, C. Poindexter, J. Chen, A. Elbashandy, M. Humphrey, et al. The fluxnet2015 dataset and the oneflux processing pipeline for eddy covariance data. *Scientific data*, 7(1):1–27, 2020.

G. Wohlfahrt, C. Irschick, B. Thalinger, L. Hörtnagl, N. Obojes, and A. Hammerle. Insights from independent evapotranspiration estimates for closing the energy balance: a grassland case study. *Vadose Zone Journal*, 9(4):1025–1033, 2010.