Reviewer comments and author responses

The structure of this author response is as follows. We first respond to all comments of reviewer 1 and afterwards to all comments of reviewer 2. The comments that have been taken from their review are underlined and in italic font. Our responses can be found below each comment in normal font.

Reviewer 1:

We would like to thank reviewer 1 for his comments. We structured them according to the following three sections: “General issues”, “Data check”, as well as “Detailed issues” and provide an answer to all of them in the following.

General issues:

1) Data access requires registration and notification. Not acceptable for an open access data journal. (Link to the BACI server at BGC-Jena works well but registration step remains an unwelcome and un-necessary barrier! As I understand, ESSD advocates open ‘one-click’ access! Why did the ESSD publishers and editors allow these particular impediments? I want to provide an anonymous review but to see the data I must supply name and email. The system sends notice to data owner (lead author in this case), who now knows my name and email address. How does that qualify as anonymous? BACI will get more data usage and better data tracking (through Thomson-Reuters doi-based data usage tracker) if they open this up.)

We sincerely apologize to the reviewer for the trouble our data portal has caused and we do understand that the registration step does not allow for a completely anonymous review. We simply have not thought about this situation and have always used our data portal to share data with the community in the past. The registration with email is a default setting of our data portal which we use exclusively to inform users in case of the availability of updated versions or potential bug fixes of respective downloaded products. We can assure that we never intended to monitor who is downloading the data explicitly, especially with respect to anonymous reviewers. Sorry again for the inconvenience. Nevertheless, the registration step does not violate open access standard and especially not the journal standards of ESSD, since on the journal webpage we found the following information regarding repository criteria as part of the author guidelines:
“Open access: The data sets have to be available free of charge and without any barriers except a usual registration to get a login free-of-charge.” (https://www.earth-system-science-data.net/for_authors/repository_criteria.html)
We therefore think that we have met these conditions with our data portal.

2) Very confusing use of the term ’diurnal’. Needs a systematic re-assessment and revision of terminology.

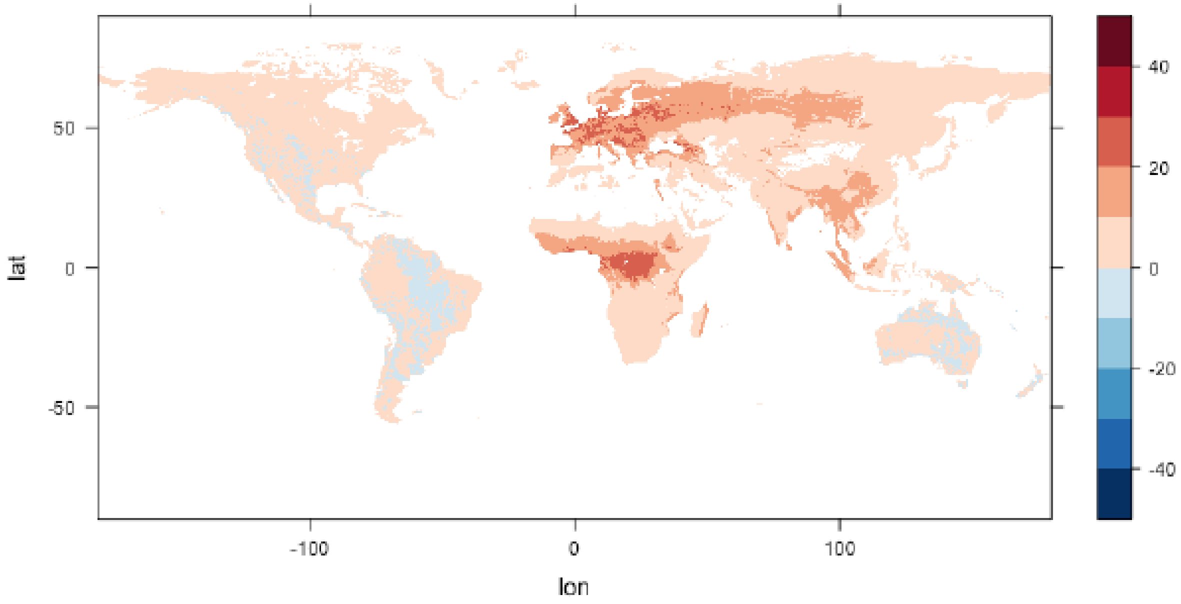
The reviewer gave more explanations for this concern in the first comment under "detailed issues" below. We therefore respond in more detail to this point there.

3) Possible data file error. If due to an ingest or plotting error by this reviewer, authors need to verify original file and contents. If confirmed, authors will need to implement a thorough check and validation across all source files. Manuscript not acceptable without one or the other outcome.

The reviewer spotted slightly negative GPP values during nighttime in the provided data products, which are then claimed as possible data file error. We have checked the files and there is no data file error. Also the plot that has been provided by the reviewer is correct. However, the reviewer has explained the issue in more detail in the next comment (Data check) and we give a detailed answer there.

Data check:

Using ncdf4 library in R I opened one of the ADC (average diel cycle) files, extracted a month and hour time slice and plotted it (below). As specified in the netCDF file, scale = 1 and offset = 0, e.g. no scaling or offset. This 4D file: GPP_halfhourly_monthlyADC.upscalingProduct_v1.array4D.720.360.2010.nc; with this float variable: GPP_halfhourly_monthlyADC (average diurnal/diel cycle?); at Hour = 20 (roughly 1200 local over North America or, if 20th position in the array rather than 20 UTC, 1000 local over Europe?); and Month = 6 (June). Apologies for the weak plot. I do not claim expertise in R but I use it semi-regularly for data processing and to open and check ESSD data files of various formats. In the plot below, I have obviously preserved correct X (lon) and Y (lat). I would not have expected negative GPP values? Please can the authors check to assure that this result represents an ingest or plotting error by me and not an error in the data files. Negative values suggest that we might have an NEE rather than GPP file? If we do have a file type or data type error, the authors then need to check and verify the full set of files?



We appreciate the reviewers effort in checking the data files and raising the issue of negative GPP values during nighttime. First of all, the plot is correct and slightly negative GPP values seem to be unreasonable at first glance. The reason for this is that the flux partitioning of NEE into GPP and terrestrial ecosystem respiration (TER) at site level can result in slightly negative values, which is an artifact of the flux partitioning method that we do not have under control. Hence, we already have negative GPP values in the training set of the machine learning model and thus, this model will also produce slightly negative values for similar environmental conditions at global scale.

In other words, consider the observed GPP as a sum of the true GPP and observational noise: GPP_obs = GPP_true + noise. Especially during night, GPP_true will be equal to zero but the noise term (depending e.g. on measurement system and device) might lead to negative GPP observations. As indicated earlier, this observational error at site level is inherent in the training set and propagates into the derived data-driven upscaling product.

Please note that the discrete color scale of the map provided by the reviewer is a bit misleading, because the light blue color that dominates in the map spans a large range of 0 to -10. In fact, we have verified that the smallest value in this map is about -1.5 (micro mole per square meter and second).

However, setting the negative values to 0 (either at site level before training or at global scale as post-processing) would generate a bias (just consider a mean value of zero with some random variations as a comparable example). We therefore agreed on keeping negative values but understand that this can cause confusion. We have therefore added a paragraph in the section "Data availability and usage notes" that points to that issue and indicated an optional post-processing step to set negative GPP values to 0 for specific applications.

Detailed issues:

Page 1 line 8: technically diurnal = daytime, nocturnal = nighttime, diel = 24 hour cycle of day plus night. Here the authors - in common with many other researchers - use 'diurnal' when in fact they mean 'diel'? They use the climatological term 'diurnal cycle' to indicate a pattern that repeats daily (e.g. light-driven carbon fixation in GPP) but NEE - which includes 24 hour respiration - should properly carry the label 'diel' cycle? These authors could and should make a more precise distinction between 'diel' and 'diurnal'. Ecologists - who might represent one of the audiences for this data set - generally use the term 'diel'. Occasionally the authors use the phrase 'diurnal courses' (e.g. page 1, line 15). This reviewer does not understand that phrase, nor will readers. Suggest 'diurnal' for light-driven processes (e.g. GPP) but 'diel cycles' or 'diel patterns' for all other 24-hour cycles. Or the authors should define their use of 'diurnal' near the top of the manuscript and then follow that usage carefully throughout. Authors can substitute diel for diurnal without changing the ADC acronym.

We thank the reviewer for introducing the term “diel” to us. We have not been aware of this word in the mentioned context. We also appreciate the detailed explanations highlighting the differences between diel, diurnal and nocturnal. However, and as mentioned by the reviewer, many other researchers also use the term “diurnal” where in fact they mean “diel” and at our institute as well as in our community (including project partners from other institutes and universities) the term “diurnal” is commonly used. We therefore stick to the term “diurnal” and follow the suggestion of the reviewer by explaining what we mean by this term at the beginning of the paper in a prominent way. In fact, we use diurnal for the full 24-hours cycle and the terms “daytime” and “nighttime” to distinguish between light-driven patterns and values during night. We also highlight that the term “diel” would also refer to the 24-hours cycle, such that other researchers that are more familiar with this term (e.g., ecologists as mentioned by the reviewer) do understand our descriptions. Furthermore, we state explicitly in the revised version that we treat the phrase “diurnal courses” as a synonym for “diurnal cycles”. Note that in the following, several comments of the reviewer refer to the difference between diel and diurnal. For these comments, our response will always point to this answer of the first detailed issue.

Page 1 line 17: ‘plain’ half-hourly flux products. Use of the word ‘plain’ here implies ordinary or simple. Not an accurate reflection of your work! Perhaps ‘full’ or ‘more extensive’?

This is a good point. We followed the suggestion of the reviewer and have replaced the word “plain” by “full” in the revised version.

Page 1 line 24: ‘underlying observations’? Confusing. Observations underlay the eddy covariance measurements? Or the eddy covariance measurements underlay land-atmosphere interactions? Better to write ‘the underlying measurements are local’

Thank you for making us aware of this confusion. We have rephrased the sentence to be more precise.

Page 2 line 2: proofed or proved?

We use the word proved in the revised manuscript.

Page 2 line 7: ‘model is being applied’. Instead, ‘model is applied’ or ‘models are applied’.

We use the phrase “model is applied” in the revised version.

Page 2 line 9: replace 'making' with 'initial' - 'do not require initial assumptions on functional relationships'

We followed the suggestion of the reviewer and replaced the word "making" with "initial".

Page 2 line 11: 'first' what? First machine learning paper? First upscaling paper? First in a series of Jung et al. 2009 papers but the reference list only shows one Jung et al. for 2009. Need clarity here!

We have written "One of the first upscaling papers" in the revised version.

Page 2 line 24: 'this paper' Which paper? Tramontana et al. 2016? This (current) ESSD paper? Vague language allows confusion here.

We rephrased the sentence and skipped the words "this paper" to avoid confusion.

Page 2 line 30: NCEP wants their full name and location specified, at least the first time you use their acronym.

We thank the reviewer for pointing us to this issue. We have inserted full name and location in this sentence.

Page 3 line 2: rises or raises?

We use the word "raises" in the revised version.

Page 3 line 4: In addition to what? Do you mean 'never the less'?

We have replaced "in addition" by "furthermore".

Page 3 line 5: a diel rather than diurnal cycle?

This refers to the general distinction between diel and diurnal that has been explained in the response to the first comment in the detailed issues section above.

Page 3 line 11: heat 'waves'? You mean heat extremes or temperature extremes? Heat waves typically have multiple day duration?

Here, we believe that the reviewer got confused why a multi-day event like a heat wave has consequences on the diurnal cycle. For heat waves, typically the daily average temperatures are reported. However, increasing average temperatures arise from larger temperatures at subdaily (e.g., half-hourly) time scales. Especially the peak temperatures around noon are usually affected and these can cause severe trouble for plants. For example, a day with an average temperature of 25° C, where all half-hourly values of the day are close to 25° C, might be better for the healthiness of plants compared to a day with the same average temperature but where a clearly colder night balances peak temperatures around 40°C at noon, which are usually causing most of the damage.

Page 3 line 22: good explanation and justification of focus on GPP

We have added a sentence for explaining and justifying the focus on GPP in the revised version.

Page 3 line 23: delete this first sentence, you don't need it.

We agree that this sentence could easily be skipped, but we like to keep it in order to have a good transition to the next paragraph that gives an overview of the structure of the paper.

Page 3 line 29: freely available after registration and certification. Your definition of 'freely available' does not match either ESSD or organisation (e.g. WDS, RDA) standards and expectations?

We do understand the irritation of the reviewer and we again apologize for the inconvenience the registration step has caused for the review process. However, we think that we do not violate ESSD standards, because on the journal webpage we found the following information regarding repository criteria as part of the author guidelines:
"Open access: The data sets have to be available free of charge and without any barriers except a usual registration to get a login free-of-charge." (https://www.earth-system-science-data.net/for_authors/repository_criteria.html)
As indicated earlier, our data portal has the registration step just to inform everybody who has downloaded the data about updates, new releases, bug fixes, etc. in a newsletter style via e-mail.

Page 4 line 1: here you clearly mean diel, not diurnal

This refers to the general distinction between diel and diurnal that has been explained in the response to the first comment in the detailed issues section above.

Page 4 line 4: only achieved temporally by eddy covariance instruments and only extended spatially by deployment of those instruments on globally-distributed towers.

We have rephrased the sentence to account for the suggestion of the reviewer.

Page 4 line 7: ‘By now, the flux towers have produced valid data for sufficient time to ... but temporal representativeness remains highly uneven’. Chu et al. 2017 indicates some risk or challenge to using flux tower data to resolve full diel cycle? Does the use of data from 222 towers somehow obviate or smooth over the factors identified by Chu et al?

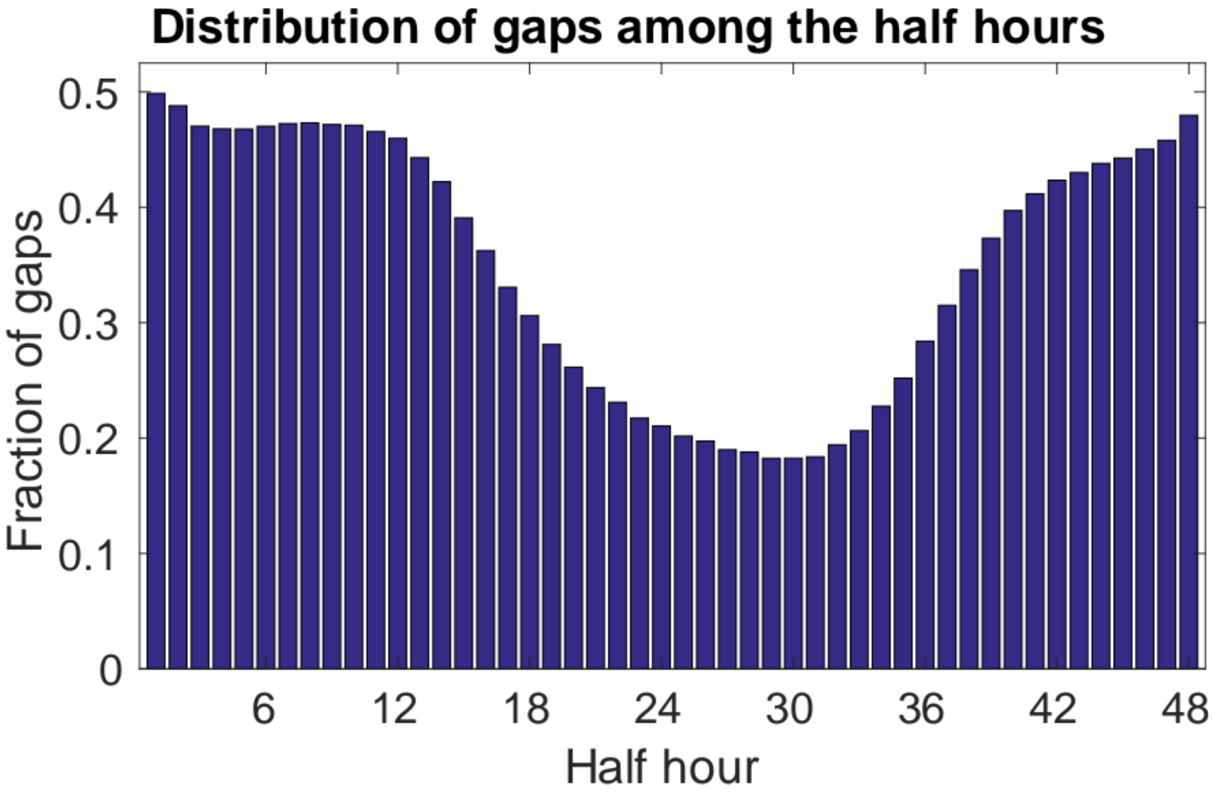
We have integrated the suggested changes of the sentence in the revised version. The paper of Chu et al. provides a general analysis for the temporal representativeness of FLUXNET with respect to years of observations. It does not indicate risks or challenges for resolving the full diel cycle, because the analysis has been carried out on a much broader time scale (monthly values to compare with climate variables from CRU). However, we think that it is not only a matter of the number of towers that are used for the upscaling but also the number of site-years or even site-days. In this respect, we have used all the available data from all the sites that have provided half-hourly flux data of sufficient quality (no gapfilled data), which is the best that we can get.

Page 4 line 10: each tower has - thanks to FLUXNET standards - instrumentation and data outputs similar enough to allow each site to serve as a valid local node in a system-wide machine learning exercise? The machine learning training is not local, it requires or assumes multiple comparable sites widely distributed?

We agree that the term “local training” is confusing here. What we actually meant is site-level training that is required prior to the global estimations. However, we have rephrased the sentence and incorporated the suggestions of the reviewer. Please note that multiple sites do not all need to be comparable in terms of environmental conditions but in terms of instruments and measurements (e.g., same units of variables, same measurement height for air temperature, etc.). With respect to the latter, all data analysis methods including machine learning techniques do require comparable measurements in order to draw reasonable conclusions.

Page 4 line 13: the authors should help us understand the frequency and impact of these gaps? 20% (my best guess) due to instrument failures, power failures, data recording failures? Can the authors assure us of random gaps over the diel operation periods or do the gaps have diurnal or nocturnal dominance? If the authors want to assure on this point, e.g. that they can ‘safely’ ignore gaps, we need a more information.

We do understand that the reviewer is concerned about the data gaps in the flux measurements. From a machine learning perspective, we can anyway only do the training on the data we have, independent of their distribution in time. Furthermore, we have decided to not include gapfilled data in order to prevent the machine learning model to adapt too much to the gapfilling method. However, we have found that within all the site-days that we have taken into account, there are roughly 35% gaps and the following figure shows their distribution among the half hours.



One can clearly observe a nighttime dominance of the gaps. For GPP, this is not a big problem, because it is assumed to be zero during night anyhow. Considering NEE, the absolute fluxes are also smaller during night compared to daytime observations. The nighttime dominance of gaps arises from less turbulence during these hours and this is an inherent problem of the measurement devices that we cannot resolve. However, it should be noted that such a biased distribution of gaps does not directly lead to a model bias as it would be the case for example for linear methods. Since we have picked random forests as a nonlinear machine learning technique, our derived models are less biased for imbalanced data because the final estimations in the leaf nodes of the decision trees are made locally in predictor space by considering mean values from samples that fall into the respective leaf node. Hence, they are independent from samples that are far away in predictor space but could potentially have higher or lower density.

In addition, we have also carried out preliminary experiments where we have only used site-days with no gaps, i.e., where all 48 half-hourly values have been available. This has then reduced the overall number of training samples massively and has clearly reduced prediction performance, most likely due to worse generalization abilities because the reduced training data did not capture all environmental conditions sufficiently well. We have included a section in the appendix of the paper that describes these issues of the data gaps and we refer to these explanations in the section indicated by the reviewer.

Page 4 line 14: Table 2 of Tramontana gets cited in this ESSD paper six times, starting here. Table 2 has only a few lines and exists in a sister (Copernicus) open access journal, so why not save the reader a few moments by simply reproducing the Table and its caption here? Several of the same co-authors, permission should occur easily? You could actually save some explanatory text here before appropriately referring the reader to more details in specific sections of Tramontana (e.g. as you do in line 22).

We followed the suggestion of the reviewer and have reproduced the table and its caption in the appendix of the revised version of the paper. We also pointed to this appendix when we mentioned the table in the manuscript.

Page 4 line 31: Here we get definition of the CRU and NCEP acronyms (even through you already used NCEP above). But CRU at Univ East Anglia UK and NCEP at Asheville or Silver Spring in USA will want their affiliations and locations listed and promoted somewhere at least once?

Thank you for bringing this up. We have added affiliations and locations in the revised manuscript.

Page 5, section 3.1 Randomized decision tree: very good description and clear figure. No external references? We assume this section therefore comes entirely from initiative and experience of authors? Reference to other work, e.g. Breiman, appears at the start of Section 3.2, implying that these authors know at least some other uses and exploration of decision tree research and literature. We can assume the authors have saved us a lot of effort with these short clear discussions sans references?

Indeed, we have quite some experience with randomized decision trees and random forests including theoretical background from machine learning lectures. Hence, the descriptions in this section are mainly based on our knowledge. However, we have included some additional references at the beginning of this section in the revised version of the paper.

Page 7 lines 11, 12: Here the authors must tell us - or must already have told us - how they use the term ‘diurnal cycles’. Do they mean light-driven daily cycles or full 24-hour diel cycles?

This refers to the general distinction between diel and diurnal that has been explained in the response to the first comment in the detailed issues section above.

Page 8 line 10: I question the use of the word ‘accurate’ here. Averages of repeated *measurements under stable environmental conditions can improve accuracy. Averages of measurements taken over - by definition - non-stationary environmental conditions do not, without additional thorough statistical treatment, qualify as more accurate? Mean wind and wind gust - both highly dependent on sensor averaging times - offer a good example? One could say more useful or less noisy, smoother - perhaps better suited for climatological averaging - but not necessarily more accurate without better definition of accuracy requirements? We almost never read about instrument accuracy or instrument noise on daily time scales?*

We agree that the word “accurate” is not precise here and we therefore follow the suggestion of the reviewer and replace “more accurate” with “less noisy” in the revised version.

Page 8 line 12: daily temporal resolution? Do the authors mean once per 24 hour time period? Two overpasses? Eight-day repeat ground tracks? I think you mean not more frequent than once per 24 period?

Thanks for pointing us to this confusing formulation. We have clarified it in the revised version.

Figures 3 and 4 clearly show a diurnal (light-driven) cycle of GPP nested within a full 24-hour diel cycle. The authors could and should use figures like these to explain their intentions and use of terms diurnal, diurnal cycle, diurnal course, diurnal pattern. Introduction of the more accurate term diel would greatly help resolve this confusion.

This refers to the general distinction between diel and diurnal that has been explained in the response to the first comment in the detailed issues section above.

In Figure 4, the vertical columns of different colors in the upper panel, derived from specific half-hour training periods in Figure 3, now go through a generalized (uniform across all 48 time periods) RDF. So the output of that RDF, the so-called second option, visuallized by the downward arrows in the lower panel, should in fact apply uniformly to all 48 time periods, not to specific time periods as indicated by the dashed arrows? E.g. the RDF outcome or output in Figure 4 should look different to the outcome shown in Figure 3?

We believe that the reviewer got confused when comparing Fig. 3 and 4. The vertical columns in Fig. 4 are not derived from specific half-hour training periods in Fig. 3. The figures denote different prediction approaches and are independent of each other. In fact, the so-called vertical columns denote input data (predictor variables) to the RDF model and arrows indicate data flow. This means that in both figures, the input data is visualized in the corresponding upper panels and this data could be the same for both prediction approaches. However, the difference between the approaches is that for the first approach, there is one regression model learned for each half hour by just using the input data from the corresponding half hour. On the other hand, the second approach in Fig. 4 denotes the scenario where only one RDF is learned using all the data of all the half hours, which requires at least one predictor variable at half-hourly resolution. In summary, both figures rather visualize the test step for a single day than the learning phase of the RDF models. We have added additional information in the figure captions to avoid confusions.

Page 9 line 16: the use of 1st diff of Rpot to distinguish rising (morning) or falling (afternoon) seems like a clever benefit of the overall Rpot approach. Did this group discover or initiate that technique? For nocturnal periods, Rpot = zero for many consecutive hours so 1st diff Rpot also zero. Here we have the basis for a definitive distinction of diurnal and nocturnal, which varies substantially with latitude and season, and which summed together give a full diel cycle. Make use of this relatively simple indicator to define diurnal vs diel?

Indeed, we came up with the idea of including the temporal derivative of Rpot to distinguish between morning and afternoon. We are not aware of others that have used this indicator for prediction purposes, but we also believe that including derivatives of predictor variables as additional predictors is a standard technique for feature generation from time series data. We therefore did not specifically search for similar applications of the derivative of Rpot. The issue of diurnal vs diel has been discussed in the first comment to the detailed issues above.

Page 10 lines 7 to 9. Here the authors contend that, in the absence of half-hour global gridded meteorological data, the high-temporal resolution of the tower-based flux and attendant micrometeorological data can prove useful in a validation of the upscaled products. But the earlier statement about ignoring temporal gaps impinges here? If those gaps amount to 20% with a distinct diel pattern - this we don’t know but presumably the authors do - then use of the flux data as a validation tool introduces additional uncertainty. Again, we don’t know the quantitative impact but, thanks to the authors, we do know that temporal gaps exist. If 2%, not a problem? If 20%, the authors need to at least assure us about random (in time) occurrence? At this point we need an answer to our earlier question: should we safely ignore those gaps or not?

This has been addressed in our response to the first question about the gaps in the data above. We therefore refer to this answer. In addition, our experiments comparing the leave-one-site-out with the leave-one-month-out (Table 2) have shown that extrapolating to a new site is a much harder problem than filling gaps (since the left out month can be seen as a long gap in a time series). There are gapfilling papers around, but as indicated in our response to the previous question about data gaps, we did not want the regression models to adapt to the selected gapfilling method. Especially for the cross-validation experiments, one can only properly validate against the non-gaps, because different gapfilling methods make different assumptions and validation against gapfilled data introduces biases in the analysis. Therefore, we only performed our experiments on the data that has been available.

Page 12 Figure 5: We need definition of the location acronyms / codes used to designate individual flux tower sites on the X axis? Otherwise the reader needs to scroll through the long list of Appendix A to find 6 specific sites? CA-Man Manitoba Black Spruce DE-Hai Hainich Germany FR-Pue Puechabon IT-Cpz Italy Castelporziano US-Goo Mississippi Goodwin Creek US-Var California Vaira Ranch. If you do it for Figure 5 you would not need to do it again for Figure 8.

We have added the definitions of the location acronyms in the caption of Fig. 5 in the revised version of the paper.

Because the authors use the phrase modeling efficiency in this section and across many of the sections that follow, we need a little more information about Nash-Sutcliffe? Somehow related to predictive skill (also called predictive performance), with values approaching 1 preferred? Derived from river forecasts but evidently much used for forecast evaluation in the NWP community? Statistically based or pattern based? Small amount of explanation would buttress the subsequent analysis sections.

We have followed the suggestion of the reviewer and added more information about the Nash-Sutcliffe modeling efficiency in the revised version of the paper.

Section 5 - very valuable. Good analysis, good figures. The identification of seasonal drought as a distinct limitation and complication seems like a very important outcome. I also like the leave-one-month-out approach, clever. Again, at the initiative of this group or did they learn that idea from some other example? They should take credit or give credit! In Figures 7, 9, 10, 11, this reviewer sees a distinct diurnal pattern of GPP nested within (and artificially centred in these plots) a repeatable 24-hour diel cycle. Presumably if we looked at NEE we would observe a different pattern, with a weaker light-synchronized distribution, extending across more or all of the full diel cycle. And just because diurnal GPP flattens or disappears in March and October (at these exclusively northern hemisphere example sites), measurable diel patterns of soil microphysics and heterotrophy and of plant biochemistry will not have disappeared?

The idea of comparing site-specific leave-one-month-out results with the global leave-one-site-out results comes from our group, at least we are not aware of any other work that has done a similar analysis. There are few papers that use the term leave-one-month-out, but this is either for indicating predictions where a single time series from a single site denotes the whole data set (local studies only taking one site into account) or for experimental protocols where one month of one site is left out but the predictions are based on a model learned from other sites and the remaining month of the selected site. Therefore, leave-one-month-out is in general used in many different ways. For us, it is a basic tool to analyze our prediction models and we think it is too trivial to explicitly take credit by highlighting this aspect within our analysis.

Reviewer 2:

We thank reviewer 2 for his comments and answer his questions in the following.

(1) Sections 1 and 6 (Introduction and Results): It is essential to provide some context regarding other competing measurement-model estimates of some of these fluxes (e.g. NOAA Carbon Tracker, etc.) and to compare the fluxes in Section 6 to these wherever appropriate.

This is a good comment. We see the 'flux tower upscaling approach' as complementary 'bottom-up' approach to the 'top-down' atmospheric inversion approach of, e.g., Carbon Tracker. This complementarity allows for cross-consistency checks of independent data streams and even for synergistic usage, for example employing the upscaled products as priors in the inversions. We followed the suggestions of the reviewer and both inserted a paragraph on this in the introduction and included a comparison of our data product with an ensemble of atmospheric inversions (including Carbon Tracker) in the results section.

(2) Section 2 and Appendix A (Data Sources): To judge the adequacy of the global land spatial coverage of the eddy covariance tower data that underpin the model development and testing, it would be essential to show a latitude-longitude map of their locations superimposed on lat-long maps of the computed variables (GPP, etc) and to comment on geographical regions where the upscaling model has not been adequately assessed using observations.

We thank the reviewer for bringing this up. We have included a corresponding map together with a small section in the revised version of the paper.

(3) Sections 6 and 8 (Data uncertainties): The half-hourly data are provided without an explicit measure of their uncertainties, which is not satisfactory if they are to be used e.g. to compare with other estimates of the computed variables, or to use e.g. as priors in a "top-down" optimal estimation of these variables using atmospheric models and CO2 mole fraction data. In this respect, while the Nash-Sutcliffe model efficiency (NSME) values given in the text help give confidence in the model, they are not directly applicable to uncertainties in the data. As a proxy for these uncertainties you could at least show the rms values of the model-observation

residuals used in the calculation of the NSME values. While these would not be applicable to the uncertainties in the poorly observed regions, they would at least provide lower limits to them.

Uncertainties for the estimations of machine learning models is a difficult topic in general and we plan to investigate this in future research as indicated in section 7. Nevertheless, we have followed the suggestion of the reviewer and included root-mean-square errors in addition to the modeling efficiencies in the revised version of the paper.

(4) Data access: Not only are registration and notification required, but finding appropriate software to facilitate the download, viewing and display took time. It would be very useful if concrete suggestions of what software works (Panoply, etc.) were given prominently in this paper and on the BACI website.

Please note that a registration prior to the data access does not violate ESSD journal standards, since on the journal webpage we found the following information regarding repository criteria as part of the author guidelines:
“Open access: The data sets have to be available free of charge and without any barriers except a usual registration to get a login free-of-charge.” (https://www.earth-system-science-data.net/for_authors/repository_criteria.html)
As suggested by the reviewer, we provide information on both useful software and data file format in the section “Data availability and usage notes” in the revised version of the paper.