GMD-2019-315

Dear GMD editor,

First, we would like to thank the associate editor and reviewers for handling our manuscript. Enclosed to this letter, a revised version of our manuscript GMD-2019-315 may be found. We believe we have addressed all comments raised by the two referees and modified the manuscript accordingly. The most important changes can be found in Sect. 2, where we have re-structured the entire content – per suggestion of reviewer 1 – to better separate the description of the data and methods used for evaluation. In addition, in Sect. 2, we now also introduce the method used for statistical testing, from which the results are further discussed in Sect. 3.

Finally, in addition to the corrections suggested by the two referees, some minor textual issues have been resolved in the revised version of the manuscript.

Below, a list of the comments per reviewer is given, together with our reply and a description of the changes in the manuscript in **bold fonts.** Please note that page and line numbers refer to the new version of the manuscript, unless explicitly mentioned. We hope that this revised version of the manuscript is eligible for publication in GMD. Attached to this letter, you may also find a marked-up version of the revised manuscript.

Brecht Martens,

On behalf of all co-authors.

RC1 (published online: 11/03/2020):

The paper investigates the skill of the land surface energy partitioning in the ERA5 reanalysis. For reference, the ERA5 skill is compared to that of its predecessor, ERA Interim (ERA-I). Skill is determined in several different ways: (1) directly vs flux tower measurements, (2) vs fluxes from water and energy balance estimates, (3) by driving the GLEAM land surface scheme with reanalysis data and validating the output vs. flux tower measurements, and, finally, (4) by driving the CLASS4GL boundary layer model with reanalysis data and validating the output vs. balloon observations. The authors find that ERA5 land surface energy partitioning is generally improved over that of ERA-I. In particular, the overestimation of the latent heat flux in ERA-I is reduced (but not eliminated) in ERA5.

The paper is of interest to GMD readers and makes an important contribution by documenting the quality of ERA5 land surface estimates. By and large, the writing and graphics are clear and concise, and the conclusions drawn from the study are supported by the results. I recommend eventual publication of the paper in GMD provided the authors address the comments below. It would be particularly helpful to include other reanalysis data in the comparison.

We thank the referee for reviewing the paper and providing us with useful comments, feedback, and corrections. Below, we give a point-to-point reply to the comments posted by the reviewer and describe the changes that have been implemented in the revised version of the manuscript. Please note that page and line numbers refer to the new version of the manuscript, unless explicitly mentioned.

Major comments (no particular order):

1. The title and introduction do not make it sufficiently clear that the turbulent fluxes investigated here are for the land only. I suggest changing the title to: "Evaluating the *land* surface energy partitioning in ERA5" and occasionally replacing the expression "surface [..] fluxes" with "land surface [..] fluxes", e.g., P3/L15, P4/L3, and probably a few more places.
   The authors agree with the referee and have updated the title as suggested. Also, the term "land-surface fluxes" is now introduced more often into the text.

   Changes in manuscript: **The title of the manuscript has been changed to "Evaluating land-surface energy partitioning in ERA5".**

2. P4/L1: Aren't there advances in land data assimilation from ERA-I to ERA5? And do these not matter for the quality of the land surface turbulent flux estimates? Land data assimilation in ERA-I and ERA5 and its impact on land surface estimates should be included briefly in the Introduction and further discussed in the Results and Discussion section.
   The authors agree with the referee that changes in the land data assimilation system, and their potential influence, were not described in sufficient detail in the manuscript.

   Changes in manuscript: **In Sect. 2.1 (P4-L1-9), we now highlight that changes in the land data assimilation system might also have an important impact on the land-surface turbulent heat fluxes. In addition, in Sect. 3.1.1 (P11-L6-10), we also hypothesise that the more advanced land data assimilation system – together with the superior land-surface model implemented in the IFS – are behind the enhanced surface energy partitioning.**

3. Section 2 is a somewhat odd mix of "methods" and "data". E.g., the FLUXNET 2015 section 2.2 includes discussion of how the climatology is derived for the computation of the standardised anomalies, but this would also apply to the other datasets (incl. ERA-I and ERA5). The entire section needs to be reorganized and be more clearly separated into "Data" and "Methods".
   The authors agree that this would improve the readability of the manuscript.

   Changes in the manuscript: The authors have re-organised Sect. 2 per suggestion of the reviewer, aiming at a better separation of the methods and data.

4. The violin plots are great visual tools, but I assume their construction involves is some fitting of the distributions. These details should be in the "Methods" section.
   The method used to construct the violin plots involves a kernel density estimation approach where the band width is calculated according to Scott (1979).

   Changes in the manuscript: The method for constructing the violin plots is now described in more detail in the revised version of the manuscript. We refer to Sect. 2.7.1 for more details; there we describe multiple aspects of the validation strategy.

5. There are gaps in the literature discussion. E.g., Draper et al (2018) is a highly relevant assessment of reanalysis estimates of land surface energy flux estimates, incl ERA-I. Draper, C. S., R. H. Reichle, and R. D. Koster (2018), Assessment of MERRA-2 Land Surface Energy Flux Estimates, Journal of Climate, 31, 671-691, doi:10.1175/JCLI-D-17-0121.1.
   The authors thank the reviewer for this suggestion.

   Changes in the manuscript: The paper by Draper et al. (2018) is now referenced in the discussion of the results. More in particular, the paper is cited in Sect. 3.2 where the biases in ERA5 are discussed and in Sect. 3.4, where the globally-averaged land-surface turbulent heat fluxes are assessed.

6. Related to (5), the present paper only investigates ERA-I and ERA5. The paper would be considerably more relevant to readers if these ECMWF products were assessed along with at least one or two other, major reanalysis products (such as MERRA-2).
   The authors thank the referee for this suggestion and fully agree that this would be an interesting analysis. However, the focus of the paper is to make a first assessment of the quality of the surface turbulent fluxes in ERA5 against different observation-based data sets. Including other reanalyses – next to ERA-I, which is there just to show improvements of ERA5 upon its predecessor – would deviate the focus from ERA5, and turn this manuscript in the direction of a comparison paper. We agree on the value of such a study in the future; yet, as we say, such a comparison feels beyond the scope of this work.

   Changes in the manuscript: No changes have been implemented in the manuscript.

7. If I understood this correctly, the GLEAM and CLASS4GL analyses work as follows: (i) use ERA-I and, separately, ERA5 to force GLEAM (or CLASS4GL), then (ii) evaluate the results against tower (or balloon) measurements. This approach depends on the assumption that GLEAM and CLASS4GL are very good models, or at least that they do not have compensating errors. If, say, errors in GLEAM were to compensate for errors in ERA-I, forcing GLEAM with a better reanalysis isn't necessarily going to

deliver better GLEAM outputs. Similarly for CLASS4GL. This is a major caveat that needs to be discussed prominently in the paper.

The authors agree that this type of analysis comes with assumptions, yet it still provides useful insights in the quality of ERA5. Unless structural errors in GLEAM and CLASS4GL are somehow (anti-)correlated, forcing either of the models with better inputs can only lead to better model output. This would therefore yield better validation results when compared against independent in situ observations, as long as it is done systematically at a large number of sites. Given the strong differences between the model concepts and assumptions of GLEAM and CLASS4GL, the authors believe that both model frameworks are sufficiently independent (i.e. their errors are sufficiently uncorrelated) to trust that a higher accuracy in the forcing data should reflect on better match to the in situ measurements after this modelling chain.

Changes in the manuscript: This caveat in the analysis is now briefly mentioned in Sect. 3.1.2 (P12-L4-11).

8. Figures A.1-A.4 are oddly placed. Either they need to be put into a proper Appendix, with discussion in the appendix as well, or they need to placed in a separate "Supplementary Information" document. As assembled, Figures A.1-A.4 are simply out of order.

   According to the guidelines of Copernicus, the authors think that Figures A1-A4 should actually be included into the manuscript as appendices:

   "Additional figures, tables, as well as technical and theoretical developments which are not critical to support the conclusion of the paper, but which provide extra detail and/or support useful for experts in the field and whose inclusion in the main text would disrupt the flow of descriptions or demonstrations may be presented as appendices. These should be labelled with capital letters: Appendix A, Appendix B etc. Equations, figures and tables should be numbered as (A1), Fig. B5 or Table C6, respectively. Please keep in mind that appendices are part of the manuscript whereas supplements (see below) are published along with the manuscript."

   "Supplementary material is reserved for items that cannot reasonably be included in the main text or as appendices. These may include short videos, very large images, maps, CIF files, as well as short computer codes such as matlab or python script."

   The text above has been exactly cited as it appears at https://www.geoscientific-model-development.net/for_authors/manuscript_preparation.html (last visit June, 1st 2020).

   Changes in the manuscript: No changes have been implemented in the manuscript, but we will follow the editorial advice on this.

9. Figures 2 and 4 are a confusing mix of dimensional metrics (MD for raw fluxes) and unit-less metrics (MD for raw Bowen ratio, and MAD and R for standardised anomalies). These figures also lack basic information about a dimensional 2nd-order metric such as MAD or RMSE for dimensional (raw or anomaly) variables. The readers are going to want to know typical MAD or RMSE values for fluxes in units of W/m2, incl. and/or excl. the seasonal cycle. I suggest revisiting the assembly of Figures 2, 4, A.1-A.4. The paper would be much easier to follow and more informative if, say, one figure includes only (dimensional) metrics computed from raw time series and another figure includes only (non-dimensional) metrics from standardised anomaly time series.

The primary reason to focus on the evaluation of anomalies is to minimise the effect of the strong seasonal cycle in the turbulent fluxes, which might mask important differences between the quality of ERA5 and ERA-I. Then, we prefer to focus on standardised anomalies to allow direct comparison of metrics from the different turbulent fluxes that typically range in a different order of magnitude. Needless to say, calculating the Mean Difference on anomaly time series is not needed, as anomalies are mean zero; hence we report the Mean Difference calculated on raw time series.

Nonetheless, we do agree with the referee that readers might be interested in the statistics calculated on raw time series as well. Note that, therefore, we report these corresponding statistics in Figures A1 and A4.

Finally, the authors believe that the content of the figures is described with sufficient detail in the captions and corresponding text to avoid confusion.

Changes in the manuscript: We better motivate our choice to evaluate standardised anomalies in Sect. 2.7.1 (P8-L26-P9-L2) of the revised version of the manuscript.

10. P10/L15-16: "Figure 4b shows that ERA5 is better at capturing..." Doesn't this invalidate the conclusion drawn from Fig 4a, that is, the evaluation of ERA-I and ERA5 through GLEAM? See also comment (7) above.
    As replied to comment #7, forcing GLEAM with better inputs can only lead to better validation statistics against in situ, when done over a large number of sites and for long-term periods. This is especially true when focussing on inferences such as anomaly correlations, and could only be rebutted when referring to e.g., long-term biases. Hence, we think the statements are correct.

    Changes in the manuscript: No changes have been implemented in the manuscript.

11. P11/L3-8: One aspect that may come into play here is that GLEAM+ERA5 is an off-line (land-only) modeling system that does not permit feedback whereas ERA5 is a coupled land-atmosphere modeling system. This may be related to a finding by Draper et al (2018): "Finally, the SH results for MERRA-Land are troubling. While MERRA-Land did have the desired reduction in the LH biases compared to MERRA (to 1 W/m2 in the global land annual average), it also had a compensating, and much larger, increase in the SH bias (up to 15W m22 in the global land average)" [beginning of p689 of Draper et al. 2018]. See also comment 5) above about the need for a better integration of the results of the present paper into the literature context.
    We fully agree with the referee that this is an important difference between both modelling approaches that may strongly affect the simulated fluxes. We also thank the reviewer for the suggested paper that fits well within the current manuscript.

    Changes in the manuscript: The fact that ERA5 is a fully-coupled system, while GLEAM is an offline land-surface model is now highlighted in the manuscript (P13-L29-31) and reference to Draper et al. (2018) is included in several contexts (Sect. 3.2 and 3.4).

12. P11/L26-P12/ L2: There is no discussion of Fig 9! In this paragraph, insert explicit references to Figs 7, 8, and 9 in the relevant place within the paragraph. E.g., reference Fig 7 in P11/L28, reference Fig 8 in P11/L33. This reveals that Fig 9 has not been discussed.
We thank the referee for picking this up.

Changes in the manuscript: Given that the conclusions for the Bowen ratio largely follow from the discussion on the turbulent fluxes, we briefly refer to Fig. 9 now in P15-L5-7.

13. In some cases the results are overstated. E.g., P12/L22: "The improvements are less clear..." suggests that there are some (hard-to-see) improvements, when in fact the results are neutral at best. P14/L3-4: The statement here is not consistent with the results of Fig 4 that show that ERA-I estimates of the sensible heat flux and Bowen ratio are better than those of ERA5.
The authors will have a detailed look at the conclusions again to soften some statements and to align the conclusions better with the results discussed in the remainder of the manuscript.

Changes in the manuscript: The manuscript will be screened for conclusions that might be too strong.

14. The tower validation results should come with some measure of statistical significance or error bars. Are the improvements, that is, the small shifts in the distributions of the metrics as shown in the violin plots, meaningful?
We agree with the reviewer that differences in quality are sometimes marginal, although often consistent. Although relying on assumptions of its own, we agree that reporting a measure of statistical significance would be useful. Therefore, we now test the differences in the anomaly correlations and the Mean Difference (MD) of the main experiments (i.e. ERA5 vs ERA-I and GLEAM vs ERA) for statistical significance (at the 5% significance level). Note that no tests were introduced for the Mean Absolute Difference (MAD), as no analytical solutions to calculate the confidence intervals are available for this metric. However, we believe that the conclusions drawn from the MAD largely follow the ones from the MD and Pearson Correlation.

Changes in the manuscript: A detailed description on the method used for statistical testing is given in Sect. 2.7.1 (P9-L11-19). Results of the statistical tests are included throughout the discussions in Sect. 3.1.1 and 3.1.2.

Minor comments:

1. P2/L13: Reichle et al. 2017 is a reference primarily for MERRA-2 land surface estimates. MERRA-2 which is a full atmospheric reanalysis, similar to ERA-I and ERA5. In this place, however, the authors are here referring to land-only reanalyses, such as ERA-Interim/Land and MERRA-Land. For the latter, Reichle et al. (2011) is a better reference. Note that there is *not* a land-only reanalysis associated with MERRA-2.
Reichle, R. H., et al. (2011), Assessment and enhancement of MERRA land surface hydrology estimates, Journal of Climate, 24, 6322-6338, doi:10.1175/JCLI-D-10-05033.1.
We thank the referee for the suggested correction.

Changes in the manuscript: The reference has been updated.

2. P2/L26-30: The text here is about very geographically limited results (southern Antarctic peninsula) or sea-ice, which is not the focus of the present paper. I suggest deleting this text or moving it further down. It confuses the reader by distracting from the focus of the paper on the global land surface turbulent fluxes.
   The main idea of this paragraph is to give a brief overview of studies that have already evaluated the quality of ERA5, irrespective of the scientific field, and to indicate that – although the number of studies is rather limited – results generally show a high quality of ERA5 compared to other datasets. We agree that these studies had a different focus than the current manuscript, but the authors believe that this short discussion is still relevant.

   Changes in the manuscript: No changes have been implemented in the manuscript.

3. P4/L19: On first reading, I completely missed the term "standardised" here and later got confused about the lack of dimensions/units in the graphics. In many papers, anomalies from the seasonal cycle are examined without standardisation, i.e., they are dimensional anomalies. There is nothing wrong per se with the standardised anomalies, but please make it clearer that you are focusing on dimensionless anomalies.
   Thanks for pointing this out.

   Changes in the manuscript: The use of standardised anomalies and the reason for this choice are now highlighted in Section 2.7.1 of the revised manuscript.

4. P10/L31: typo: "we should emphasis" –> "we should emphasise"
   We thank the referee for their detailed look at the paper and picking this up.

   Changes in the manuscript: The typo has been corrected.

5. P12/L16+L20: The numbers referenced here contradict the numbers in the graphic.
   We thank the referee for picking this up, the numbers were indeed incorrectly reported.

   Changes in the manuscript: This has been corrected.

6. Caption Fig 2: Replace "For MD, the distribution of beta…" with "The distribution of the MD of beta…"???
   We thank the referee for the suggestion.

   Changes in the manuscript: The caption has been updated per suggestion of the reviewer.

7. Caption Fig 7: "…between the absolute bias in ERA5 and ERA-I;"??? Maybe I'm misunderstanding this, but I think the bottom panel shows abs(bias(ERA-I)) minus abs(bias(ERA5), that is, the sign of the abs bias difference is different from what the caption suggests.
   We thank the reviewer for pointing this out. The caption should indeed read: "The bottom map represents the difference between the absolute bias in ERA-I and ERA5; hence, green colors represent lower bias in ERA5 than in ERA-I."

   Changes in the manuscript: The caption has been corrected.

8. Caption Fig 10: "…versus ERA-I (squares)"??? Should this read "…versus ERA-I (triangles)"???

We thank the referee again for his detailed look at the figures, the symbols were indeed wrongly referenced in the caption.

Changes in the manuscript: The caption has been updated per suggestion of the reviewer.

9. Figure 11 needs units on the colorbars. I also suggest making the graphic bigger so it can be read more clearly in a hard copy for the next round of reviews.
   The authors agree that the figures were too small and that units need to be included on the colorbar.

   Changes in the manuscript: The size of the figure has been increased and units have been added.

RC2 (published online: 24/03/2020):

General comments:
This is a very interesting and useful study of the representation of the land surface energy budget in European global atmospheric reanalyses. A large number of diverse in situ observations are used to benchmark several simulations at a global scale. Overall, the paper is well written, apart from the mixing of results and discussion/interpretation. Quality of some Figures could be improved. Colour scales are sometimes confusing as "green" tends to look blue. Violin plots are useful but do not provide a point by point comparison. Could all the corresponding scatter plots be given in a Supplement? A discussion on the impact of land cover is lacking. Recommendation: minor revisions.
We would like to thank the referee for reviewing the paper and giving some interesting comments and feedback. Below, we give a point-to-point reply to the comments posted by the reviewer and list the changes that will be implemented in the manuscript.
Regarding the scatterplots, the authors believe that the violin plots together with the discussion of the results should give a sufficiently detailed understanding of the results and prefer not to include these additional figures.

Particular comments:
1. P. 1, Title: should be more specific. For example: “Evaluating the land surface energy partitioning in European global atmospheric reanalyses”.
   We believe the reviewer means 'less specific', maybe. The authors prefer to emphasise that the focus of the paper is on the evaluation of the state-of-the art ERA5 reanalysis, rather than European reanalyses in general. Note that ERA-Interim only serves as a benchmark here to show the improvements. In addition, we believe mentioning the model/dataset in the title a requirement at GMD. However, note, we did slightly modify the title to highlight that the manuscripts focusses on land-surface fluxes only.

   Changes in manuscript: The title of the manuscript has been changed to “Evaluating land-surface energy partitioning in ERA5”.

2. P. 3, L. 1-2 (“perform better than ERA5”): any reference on this?
   As this sentence builds upon the previous sentence, the reference supporting this statement is Urraca et al. (2018).

Changes in manuscript: The reference has been cited again to support this statement (P2-L30-35).

3. P. 4, L.7: I would be more specific. For example: "the more evolved HTESSEL land surface model in ERA5".
   We agree with the suggestion of the referee.

   Changes in manuscript: The sentence has been updated per suggestion of the referee.

4. P. 4, L. 19: could you explain how these anomalies are defined and calculated?
   The standardised anomalies are simply calculated by subtracting from the raw time series (1) the climatological expectation (i.e. the mean of the variable under consideration over at least 5 years for a certain time step) and (2) dividing by the standard deviation of that climatological expectation.

   Changes in manuscript: The calculation of the standardised anomalies is now described in Sect. 2.7.1 of the revised version of the manuscript.

5. P. 5, L. 1-3: It seems that a key issue was not addressed. Land cover type in ERA5 may not correspond to the tower's one. E.g. a grassland Fluxnet site may be located in an ERA5 grid cell mainly covered by forests. How did you handle this?
   We agree with the referee that this is an issue in the in situ evaluation strategy, as it is always the case in comparisons of grid cell values to in situ data. The mismatch in spatial footprint between the in situ measurement and the grid cell of the models typically leads to an overestimation of the model error, and is often referred to as the representativeness error. In this study, we did not apply any filtering to maximise the representativeness of the in situ measurements – and hence to minimise this representativeness error. However, we do agree that this issue should be explicitly mentioned in the discussion of the results.

   Changes in manuscript: This issue is explicitly highlighted in the revised version of the manuscript now (P8-L24-25, P12-L1-2, and P13-L23-25).

6. P. 6, L. 5: could you define "non-overlapping moving windows"?
   Non-overlapping (moving average) windows refer to the fact that the time windows used for calculating the averaged quantities do not intersect (e.g. Dehghani et al., 2019). Moving windows are commonly calculated for all data points of a time series, i.e. for a simple example with a window length of 5 months, the centered moving average (window) of March contains data from January, February, March, April and May, whereas the moving average centered in April is based on nearly the same data, except ranging from February to June (i.e., adjacent moving averages share some data, and are thus 'overlapping'). In this given example, the 'next' non-overlapping window would be centered in August (June–October), as the time window used for the centered average of March and this time window do not intersect.

   Changes in manuscript: This specific processing step is now better described in Sect. 2.3.1 (P6-L6-11).

7. P. 6, L. 12 (G as a fixed fraction of Rn): Could this explain the poor scores obtained for sensible heat flux in Figure 8? The soil heat flux is related to soil properties and can be influenced by sensible heat exchange with rainwater (e.g. Zhang et al. https://doi.org/10.5194/acp-19-5005-2019).
   As was described in the original manuscript at P12-L3-7, the results for the sensible heat flux should indeed be interpreted with care as they are (among others) affected by this specific assumption. However, the magnitude of the ground heat flux at daily scales is often substantially smaller than that of the other fluxes, so the authors expect only a minor impact of this assumption on the analyses.
   In addition, the approximation used in this study to calculate the ground heat flux is not uncommon, as the magnitude of the ground heat flux typically scales with net radiation (see e.g. Kustas and Daughtry, 1990; Santanello and Friedl, 2003). Also note that, although we do not explicitly account for the effect of soil properties on the ground heat flux, we do account for the land cover type, as described by Miralles et al. (2011) and Martens et al. (2017) – see also the response to the following comment.

   Changes in manuscript: The calculation of the ground heat flux is now described in detail in the Sect. 2.3.3 of the revised manuscript (P6-L17-26).

8. P. 6, L. 12 ("land cover"): which land cover? Is it the land cover used in the model?
   The ground heat flux is calculated in this paper as described by Miralles et al. (2011) and Martens et al. (2017). In essence, the ground heat flux is calculated as a fixed fraction of net radiation, depending on the sub-pixel land cover heterogeneity. The latter is parameterised using the MOD44B Vegetation Continuous Fields product, describing each grid cell as a fraction of tall vegetation (e.g. trees), low vegetation (e.g. grass), and bare soil. For the fraction of tall vegetation, the ground heat flux is 10% of the net radiation, while for the fractions of low vegetation and bare soil the corresponding percentages are 20% and 35% (Miralles et al., 2011). In the end, the fraction of net radiation assumed to be converted into the ground heat flux at a given pixel is the weighted average of the former percentages considering the fractional land covers (MOD44B).

   Changes in manuscript: The calculation of the ground heat flux is now described in detail in the Sect. 2.3.3 of the revised manuscript (P6-L17-26).

9. P. 8, L. 3: Is there an impact of the land cover type?
   The authors have tried to relate improvements/degradations from ERA-Interim to ERA5 to different ancillary data sets like land cover, elevation, and climate, but no conclusive results were obtained. Given the uncertainties in such an analysis, the authors have chosen not to further discuss these results.

   Changes in manuscript: No changes have been implemented in the manuscript.

10. P. 8, L. 17: Seasonality removal should be described is chapter 2.
    As replied to comment #4, the procedure has been described in more detail in the revised version of the paper.

    Changes in manuscript: The calculation of the standardised anomalies is now described in Sect. 2.7.1 of the revised manuscript.

11. P. 8, L. 28: what about Fluxnet site distribution in terms of vegetation types?
    As replied to comment #8, the FLUXNET sites are indeed not well-distributed across vegetation types and climate.

    Changes in manuscript: The fact that FLUXNET sites are not uniformly-distributed across the global land surface is now mentioned several times in the manuscript (e.g. P4-L26-29, P11-L1-4, P18-L6-9).

Editorial comments (Figures):

1. Figure 1: Sites cannot be easily spotted. Colors of dots and background should be changed. What about land cover types? Format of the subfigure on the right should be consistent with format of Figure 3.
   We agree with the reviewer that the details on the figure are hard to read.

   Changes in manuscript: The figure has been updated to increase the readability, by including zoomed-in panels on regions of interest.

2. Figure 6: green or blue?
   We agree with the reviewer that the colours might be confusing for some readers, but we think this is comment is purely linguistic and that the figures are clear.

   Changes in manuscript: No changes have been implemented in the manuscript.

3. Figure 10 (top subfigures): meaning of the red lines? These metrics are a bit obscure. Why not comparing scatterplots of ABL heights?
   We appreciate the suggestion. As described in the caption, the solid lines represent the median and inter-quartile range (green for ERA5 and red for ERA-Interim). The reason why diurnal changes in temperature, humidity and ABL height are compared – as opposed to afternoon temperature, humidity and ABL height – is to reduce the influence of errors in the morning initial conditions. This in addition increases the comparability of the results, as discussed by Wouters et al. (2019).

   Changes in manuscript: A legend has been added to the figure.

4. Figure 11: Not readable. Difference figures should be expanded. Green or blue?
   The authors agree that the figures were too small.

   Changes in manuscript: The size of the figure has been increased and units have been added.

References:

1. Dehghani, A. et al.: A Quantitative Comparison of Overlapping and Non-Overlapping Sliding Windows for Human Activity Recognition Using Inertial Sensors. Sensors, 19, 5026, 2019.
2. Kustas, W. P. and Daughtry, C. S. T.: Estimation of the soil heat flux/net radiation ratio from spectral data, Agric. For. Meteorol., 49, 205–233, 1990.
3. Martens, B. et al.: GLEAM v3: Satellite-based land evaporation and root-zone soil moisture, Geosc. Model Dev., 10, 1903–1925, 2017.

4. Miralles, D.G. et al.: Global land-surface evaporation estimated from satellite-based observations, Hydrol. Earth Sys. Sc., 15, 453–369, 2011.
5. Santanello, J. A. and Friedl, M. A.: Diurnal Covariation in Soil Heat Flux and Net Radiation, J. Appl. Meteor., 42, 851–862, 2003.
6. Wouters H., et al.: Atmospheric boundary layer dynamics from balloon soundings worldwide: CLASS4GL v1.0, Geosc. Model Dev., 12, 2139–2153, 2019.

Evaluating the ~~surface~~ land-surface energy partitioning in ERA5

Brecht Martens[1], Dominik L. Schumacher[1], Hendrik Wouters[1,3], Joaquín Muñoz-Sabater[2], Niko E.C. Verhoest[1], and Diego G. Miralles[1]

[1]Laboratory of Hydrology and Water Management – Ghent University, Coupure links 653, 9000 Ghent, Belgium
[2]European Centre for Medium-Range Weather Forecasts (ECMWF), Shinfield Park, Reading, RG2 9AX, United Kingdom
[3]Flemish Institute for Technological Research – Environmental Modelling Unit, Boeretang 200, 2400 Mol, Belgium

Correspondence: Brecht Martens (brecht.martens@ugent.be)

Abstract.

Climate reanalyses provide a plethora of global atmospheric and surface parameters in a consistent manner over multi-decadal time scales~~and at the global scale~~. Hence, they are widely-used in many fields, and an in-depth evaluation of the different variables provided by reanalyses is a necessary means to provide feedback on the quality ~~and potential improvements~~ to their users and the operational centers producing these data sets, and to help guiding their development. Recently, the European Centre for Medium Range Weather Forecast (ECMWF) released ~~its~~ the new state-of-the-art climate reanalysis ERA5, following up on its popular predecessor ERA-Interim. Different sets of variables from ERA5 were already evaluated in a handful of studies, but so far, the quality of ~~surface~~ land-surface energy partitioning has not been assessed yet. Here, we ~~assess the quality of~~ evaluate the surface energy partitioning over land in ~~ERA5 over land by means of evaluating~~ ERA5, and concentrate on the appraisal of the surface latent heat flux, surface sensible heat flux, and Bowen ratio against different reference data sets and using different modelling tools~~, and compare it to the quality of ERA-Interim~~. Most of our analyses point towards a better quality of surface energy partitioning in ERA5 than in ERA-Interim, which ~~can probably~~ may be attributed to a better representation of land-surface processes in ERA5, ~~but~~ and certainly to the better quality of near-surface meteorological variables~~, as evidenced by our analysis~~. One of the key shortcomings of the reanalyses identified in our study is the overestimation of the surface latent heat flux over land, which – although substantially lower than in ERA-Interim – still remains in ERA5. Overall, our results indicate the high quality of the surface turbulent fluxes from ERA5 and ~~its general improvements as compared to~~ the general improvement upon ERA-Interim, thereby ~~supporting~~ endorsing the efforts of ECMWF to improve their climate reanalysis and to provide useful data to many scientific and operational fields.

1 Introduction

The partitioning of available energy at the land surface into sensible and latent heat exerts a strong control on atmospheric boundary layer (ABL) dynamics and informs on the coupling strength between land and atmosphere. It translates variations in

the state of the land surface (e.g. soil moisture) into changes in the state of the atmosphere (e.g. cloud formation, near-surface air temperature, and the ~~atmospheric boundary layer~~ ABL height), both at local and remote locations (Teuling et al., 2017; Miralles et al., 2016; Guillod et al., 2015; Taylor et al., 2012; Seneviratne et al., 2010). Hence, surface energy partitioning is a crucial process in the occurrence and development of extreme events such as droughts and heatwaves (Miralles et al., 2018, 2014; Teuling et al., 2010; Seneviratne et al., 2006). An accurate representation of the processes involved in this partitioning in land-surface models is thus essential to advance our understanding of past variations in climate, and leverage our abilities to predict future climate and its impacts on our biosphere (Berg and Sheffield, 2018; Dirmeyer et al., 2017).

Climate reanalyses are data sets describing the past and present state of our climate system and are derived using coupled numerical models in which a vast amount of observations is ingested through a state-of-the-art data assimilation system. They typically cover multi-decadal periods and are produced using a constant model set-up and data assimilation framework (often referred to as the Integrated Forecast System, IFS), resulting in consistent data sets describing the recent state of the atmosphere, ocean, and land surface at the global scale. Therefore, reanalyses are widely used to study past climate, to derive long-term records of essential climate variables, to initialise climate or Earth system models, or to force land-surface models offline. The latter ~~approach typically results~~ may result in higher-resolution specialised land-surface reanalyses ~~(Muñoz Sabater, 2019; Albergel et al., 2018; Reichle et al., 2017; Balsamo et al., 2015)~~(Muñoz Sabater, 2019; Albergel et al., 2018; B. During the last decade, several climate reanalyses have been produced, such as the Modern-Era Retrospective analysis for Research and Applications version 2 (MERRA-2; Gelaro et al. (2017)) from the National Aeronautics and Space Agency (NASA), the Japanese 55-year ReAnalysis (JRA-55; Kobayashi et al. (2015)) from the Japanse Meteorological Agency (JMA), and the ECMWF ReAnalysis Interim (ERA-I; Dee et al. (2011)) from the European Centre for Medium Range Weather Forecast (ECMWF). Recently, ECMWF released ERA5 ~~(Hersbach et al., 2020, 2018)~~(Hersbach et al., 2020), a new global climate reanalysis currently spanning the period 1979–present, which serves as the successor of ERA-I. ERA5 is produced using an enhanced modelling and data assimilation framework and it benefits from the assimilation of a significantly higher number of improved observations compared to ERA-I. In addition, the archive will soon cover the period 1950–present and data will become available with a latency of 2 to 4 days. Finally, data are provided at a ~~better~~ higher spatial (31 km vs. 80 km) and temporal (hourly vs. 3-hourly) resolution than ERA-I. Note that in case of ERA-I, the 3-hourly resolution can, in fact, only be obtained by combining ~~forecast~~ forecast and analysis steps (Dee et al., 2011).

The number of studies evaluating the quality of different variables from ERA5 is still limited~~; yet~~. Yet, results generally point to improvements ~~as compared to~~ upon its predecessor and to a better quality than other existing reanalyses for various surface and atmospheric variables (Hersbach et al., 2020). Tetzner and Thomas (2019), for instance, evaluated several meteorological parameters from ERA5 and ERA-I over the southern Antarctic peninsula, and concluded that the better spatio-temporal resolution at which physical processes are ~~re-solved~~ resolved in ERA5 positively affects the representation of these variables. These results ~~are~~ were confirmed by Wang et al. (2019), who compared the quality of a similar set of near-surface meteorological parameters from ERA5 and ERA-I by means of in situ validation and a modelling exercise where a thermodynamic sea-ice model was forced with reanalysis data over the Arctic ice sheet. Jiang et al. (2019) and Urraca et al. (2018), on the other hand, validated ERA5 radiation components against in situ measurements and compared their quality to other reanalyses, ground-

based observations, and satellite data. Although a small positive bias still remains in ERA5 surface irradiance according to the authors – mainly due to errors in the simulation of cloud properties – ~~the bias~~ it is significantly lower than in ERA-I and MERRA-2, especially at inland ~~stations~~locations (Urraca et al., 2018). However, in more complex terrain such as mountainous or coastal regions, high-resolution regional-scale reanalyses such as the COnsortium for Small-scale MOdeling (COSMO) REAnalysis version 6 (COSMO-REA6) from the German weather service, perform better than ERA5 (Urraca et al., 2018). Also surface wind fields have been shown to be accurately represented in ERA5 (Olauson, 2018), mainly as a result of the ~~higher spatial resolution~~ relatively high spatial resolution at which physical processes are resolved. Other studies have focused on the validation of vertical profiles of atmospheric properties such as humidity and temperature, typically revealing that the representation of these fields is better in ERA5 than in various other data sets, including its predecessor ERA-I (e.g. Brunamonti et al., 2019; Graham et al., 2019; Zhang and Cai, 2019). ~~Finally, Albergel et al. (2018)~~Indirect evaluations of variables derived from ERA5 have also been performed through different hydrological modelling studies: Albergel et al. (2018), for instance, compared the quality of ERA-I and ERA5 by forcing the Interactions between Soil, Biosphere, and Atmosphere (ISBA) land-surface model with meteorological parameters derived from both reanalyses and comparing the simulated land-surface parameters from ISBA to independent data from satellite observations and in situ measurements. Based on their study, Albergel et al. (2018) concluded that ~~the model forced~~ forcing the model with ERA5 ~~performs consistently better – especially in terms of simulated~~ surface meteorology yielded consistently better estimates of hydrological states and fluxes~~– than the model forced with~~ . Finally, Tarek et al. (2020) forced two hydrological models for a large number of catchments across the Continental United States (CONUS) to show the improvements of precipitation and near-surface air temperature from ERA5 upon ERA-I.

Despite the importance of an accurate representation of the processes involved in the surface energy partitioning, at present and to the authors best knowledge, no ~~studies have~~ study has directly evaluated the partitioning of energy in ERA5 into the two major surface turbulent fluxes over land (i.e. the surface sensible and ~~surface latent heat flux~~latent heat fluxes). As surface energy partitioning acts as a nexus between the land surface and atmosphere, such an analysis might provide useful insights to further improve the modelling of this coupled system, and to advance the quality of future reanalyses. Therefore, the objective of this study is to evaluate the surface turbulent fluxes (and their ratio; i.e. the Bowen ratio) from ERA5 for the period ~~1983–2014~~ 1983–2018 at different spatio-temporal resolutions. Several experiments are conducted using various observational data sets and modelling tools to evaluate the spatial and temporal variability of ~~these variables~~ the turbulent fluxes at different scales, ranging from point to catchment-scale and sub-daily to yearly scales. The paper is organised as follows: in Sect. 2 we describe the experimental set-up and the data sets used in this study, and provide a brief overview of the key differences between ERA-I and ERA5. In Sect. 3 we describe the results of our experiments and discuss the quality of surface energy partitioning in both reanalyses; concluding remarks are summarised in Sect. 4.

2 ~~Methods~~ Data and ~~data~~methods

2.1 ~~ERA5 and ERA-I~~ Reanalyses data

ERA5 is the latest state-of-the-art reanalysis produced at ECMWF ~~(Hersbach et al., 2020, 2018)~~(Hersbach et al., 2020), replacing the widely-used ERA-I~~reanalysis~~ (Dee et al., 2011). A first segment of the data set, covering the period 2010–2016, was released early 2017, about a decade after the successful release of ERA-I. Compared to ERA-I, which uses IFS cycle 31r1, ERA5 is produced using an improved version of ECMWF's modelling and data assimilation system (IFS cycle 41r2) and ingests information from a substantially larger volume of improved observations, resulting in a high-quality reanalysis of global atmospheric, oceanic, and land-surface fields at hourly time steps, 137 vertical pressure levels, and ~~at horizontal spatial~~ horizontal resolution of approximately 31 km. Several ~~of the changes relative to~~ advancements upon ERA-I are expected to affect the surface energy partitioning in ERA5 ~~(Hersbach et al., 2020, 2018)~~(Hersbach et al., 2020), including (1) a better forcing of solar irradiance, greenhouse gases, and stratospheric sulphate aerosols, which affect the available energy at the surface that strongly drives the turbulent fluxes, (2) a substantially ~~better~~ higher spatial resolution, allowing for a more realistic representation of surface-atmosphere interactions in complex terrain such as mountainous or coastal regions, ~~and~~ (3) a ~~better~~ more advanced land-surface model, namely the Hydrology Tiled ECMWF Scheme for Surface Exchanges over Land (H-TESSEL), which has ~~already been proven to enhance the quality of simulated~~ a demonstrated high skill to simulate surface turbulent heat fluxes in offline experiments (Balsamo et al., 2015; Albergel et al., 2012; Balsamo et al., 2008), (4) improvements in the atmospheric data assimilation component, mainly affecting the atmospheric forcing of the turbulent fluxes, and (5) an evolved land data assimilation system ingesting both snow and soil moisture observations into the land-surface model of the IFS, improving the land-surface control on the turbulent fluxes.

Here, the surface sensible heat flux, surface latent heat flux, and Bowen ratio derived from both ERA5 and ERA-I are evaluated for the period 1983–2018 (i.e. the period ~~covered with reference data~~ for which reference data are available; see Sects. 2.2 ~~and 2.3) and the global domain~~–2.4) and across the global land surface. Next to the turbulent fluxes and the Bowen ratio, precipitation, 2-meter air temperature, and surface radiation components (from which the surface net radiation is calculated) are processed. These variables are used to disentangle the role of ~~improved atmospheric data and the better~~ the improved atmospheric forcing versus the more evolved land-surface model ~~on the quality of~~ in ERA5~~surface fluxes.~~. All variables are downloaded at their native spatio-temporal resolutions and temporally aggregated to both 3-hourly and daily time intervals.

2.2 ~~FLUXNET 2015~~Eddy-covariance data

In situ ~~validation data for~~ eddy-covariance data of the turbulent fluxes (i.e. ~~measured sensible and latent heat fluxes using the eddy-covariance technique~~the land-surface latent flux and land-surface sensible heat flux) are obtained from the FLUXNET 2015 synthesis data set ~~providing data for~~ covering the period 1991–2014. The ~~data set is~~ fluxes are processed as in Martens et al. (2017), including (1) masking of rainy intervals at hourly time steps to remove unreliable measurements due to wet sensors, (2) removing gap-filled data records, and (3) aggregating to both 3-hourly and daily temporal resolutions. Note that for the temporal aggregation, 20% of the higher resolution data within the interval are allowed to be missing. Aiming at the calcula-

tion of robust validation statistics, only sites with at least 365 daily records (i.e. at least one full year of data) after masking are retained, resulting in a ~~validation data set of turbulent fluxes from~~ sample of 143 quality-checked eddy-covariance sites (Fig. 1). About 50% of these selected sites have a record length of more than 10 years, with a maximum of 21 years. Note that the same set of towers is used in the sub-daily (i.e. 3-hourly) and daily evaluations of the turbulent fluxes, making the validation metrics between experiments ~~comparable. As the temporal variability of the turbulent fluxes is strongly influenced by the seasonal cycle of its main drivers, the performance of the land-surface scheme in response to anomalous weather conditions (i.e. with respect to the seasonal cycle) might be masked when raw time series are analysed. As such, most of the analyses in this study are based on standardised anomalies to better evaluate the skill of ERA5 in capturing the effect of specific meteorological conditions on surface energy partitioning. To calculate the climatology, only FLUXNET sites with a minimum record length of five years are considered, resulting in 77~~ inter-comparable. As shown in Fig. 1, eddy-covariance ~~towers when anomalies are used for validation (Fig. 1). Note that about 50% of these selected sites has a record length of more than 10 years, with a maximum of 21 years.~~ sites are not uniformly distributed across the global land surface and hydro-climatic regimes are not equally re-presented within the data set. As most sites are located in the CONUS and Europe, warm and humid regions such as the tropics are only poorly covered. Hence, results presented in this paper should be interpreted with the shortcomings of the FLUXNET 2015 data set in mind, as further discussed in Sect. 3.

~~Location of the selected eddy-covariance sites. Sites with a record length of less than 5 years (i.e. where no anomalies are calculated) are plotted in green and sites with a record length of more than 5 years (i.e. where anomalies are calculated) are plotted in yellow. Sites where measurements of meteorological data are also available are indicated with a diamond. The background provides information on the climatological (1983–2018) annual temperature and precipitation derived from ERA5.~~

~~The~~ The daily Bowen ratio at each eddy-covariance site is calculated ~~at daily temporal resolution~~ as the ratio of the ~~surface~~ land-surface sensible heat flux and the ~~surface~~ land-surface latent heat flux. The Bowen ratio might be highly unstable when the turbulent fluxes are small compared to the measurement error of the eddy-covariance system, even at the daily temporal resolution. Therefore, outliers in the in situ time series of the Bowen ratio are masked by removing records outside the following window: $[q_{25} - 1.5(q_{75} - q_{25}); q_{75} + 1.5(q_{75} - q_{25})]$, where q_{75} and q_{25} are the 75% and 25% quantiles of the Bowen ratio time series, respectively (Martens et al., 2016). ~~Next~~

Finally, next to the turbulent fluxes, measurements of surface net radiation, near-surface air temperature, and precipitation at the eddy-covariance sites are processed ~~in a similar way~~as well using a similar approach as for the turbulent fluxes, except for the masking of rainy intervals. ~~These~~ As these variables are typically not recorded at each eddy-covariance site~~and are therefore~~ , they are only available at 83 sites in total.

~~For each eddy-covariance site in the validation data set, the variables from the corresponding ERA5 and ERA-I grid cell are extracted at their native spatial resolution and as both 3-hourly and daily (temporal) aggregates. The Bowen ratio from both reanalyses is calculated and processed in a similar manner as for the in situ data, including the removal of outliers and only at the daily resolution. Eddy-covariance sites located within the same grid cell of the reanalysis are treated separately in the validation to avoid potential problems resulting from merging sensors with different absolute values and gaps in their record (Martens et al., 2017). In summary, the FLUXNET 2015 data set is used here to evaluate the surface energy partitioning~~

~~– and some key meteorological drivers of the turbulent fluxes – in ERA5 and ERA-I at both 3 hourly and daily temporal resolution for the period 1991–2014.~~

2.3 Catchment water and energy-balance data

~~Whenever~~ If changes in water storage ~~can be~~ are neglected, the catchment-scale latent heat flux can be calculated as precipitation minus river discharge; both averaged over a sufficiently long time period (Miralles et al., 2016; Liu et al., 2014; Wang and Dickinson, 2012; Miralles et al., 2011; Vinukollu et al., 2011)~~and by~~ . By taking into account the latent heat of vaporisation and the density of water:

$$\lambda\rho E = \lambda\rho(P - Q), \tag{1}$$

where λ is the latent heat of vaporisation of water (assumed to be constant; $2260{\cdot}10^3$ J kg^{-1}), ρ is the density of liquid water (assumed to be constant; 1000 kg m^{-3}), E is terrestrial evaporation (m s^{-1}), $\lambda\rho E$ is the ~~surface~~ land-surface latent heat flux (W m^{-2}), P is precipitation rate (m s^{-1}), and Q is the river discharge (m s^{-1}). The assumption that changes in catchment water storage can be ignored requires the consideration of a sufficiently long period ~~as~~ compared to the concentration time of the catchment; ~~as in Miralles et al. (2016)~~often, a yearly aggregation period ~~will be considered here~~is considered to be sufficient (see e.g. Miralles et al., 2016).

A similar reasoning as for the catchment mass balance can be made in terms of energy balance: when changes in energy storage can be neglected, the energy balance at the catchment implies that the ~~surface~~ land-surface sensible heat flux can be calculated as the difference between surface net radiation and the sum of ground and latent heat fluxes:

$$H = R_\mathrm{n} - (G + \lambda\rho E), \tag{2}$$

where H is the surface sensible heat flux (W m^{-2}), R_n is the surface net radiation (W m^{-2}), and G is the ground heat flux (W m^{-2}). Combining Eqs. 1 and 2 thus provides a means to evaluate the long-term average catchment-scale Bowen ratio, derived from surface net radiation, ground heat flux, precipitation, and river discharge as:

$$\beta = \frac{(R_\mathrm{n} - G)}{\lambda\rho(P - Q)} - 1, \tag{3}$$

where β (–) is the Bowen ratio.

In this study, Eqs 1–3 are used in combination with an observational data set of river discharge covering the period 1983–2014 to derive an annual ~~validation~~ benchmarking data set of ~~the~~ turbulent fluxes and ~~the~~ Bowen ratio at the catchment scale.

2.3.1 Discharge

Discharge measurements are obtained from the Global Runoff Data Centre (GRDC), providing data for nearly 4000 catchments with a daily or monthly temporal resolution. As in Miralles et al. (2016), records with data artifacts are first removed based on an

exhaustive visual screening, and only catchments with an area ~~between~~ larger than 2500 km^2 ~~and 100·10^3 km^2~~ are considered. In addition, only catchments with a gridded area (on a regular 0.25° latitude–longitude grid) deviating less than 20% from the area reported by GRDC are retained. If measurements are recorded at multiple locations and thus for different drainage areas (particularly in Central Europe), measurements further downstream are favoured. By doing so, catchments are selected without any spatial overlap (due to possible sub-catchments measured upstream). After this initial filtering, data available at the daily scale are first aggregated to monthly values, given that at least 25 days per month are present. To reduce the impact of e.g. human disturbances ~~(e.g.~~ such as large-scale groundwater pumping or ~~construction works in the catchment) on our analysis~~regulations of river flow, non-overlapping~~moving windows~~, centered moving averages containing monthly data of 15 years are calculated ~~over the time series. Any catchments~~ as described in Dehghani et al. (2019). Any catchment for which the average of a window ~~exceeds~~ is exceeded more than three ~~of its~~ standard deviations by the mean of the subsequent window are discarded to remove catchments where obvious ~~changes in mean river discharge occur over~~ disturbances occur during the study period. Finally, monthly averages are aggregated to annual averages, conditioning on at least 10 months per year being present.

2.3.2 Atmospheric forcing

Surface net radiation and precipitation to derive catchment-scale validation data for the turbulent fluxes and the Bowen ratio using Eqs. 1–3 are taken from the respective reanalysis in order to mainly evaluate the effect of the land-surface ~~control~~ scheme in the IFS on the surface energy partitioning, rather than the combined effect of the atmospheric and ~~land surface model.~~ land-surface model. Therefore, the reanalyses data (Section 2.1) are temporally aggregated to the annual resolution and spatially aggregated to the scale of the catchments.

2.3.3 Ground heat flux

The ground heat flux is calculated as a fixed fraction of the surface net radiation depending on the land cover as in Martens et al. (2017, 2016) ~~, using~~ and Miralles et al. (2011). The land cover is parameterised by the Global Vegetation Continuous Fields product (MOD44B v6; Dimiceli et al. (2015)) derived from measurements of the MODerate-resolution Imaging Spectroradiometer (MODIS)~~, and the open water product from Tuanmu and Jetz (2014). Alltogether, the data set described in this section is~~ . Hence, each grid cell is covered by a certain fraction of tall vegetation (e.g. forests), low vegetation (e.g. grasslands), and bare soil. For the fraction of tall vegetation, the ground heat flux is assumed to be 10% of the net radiation, while for the fractions of low vegetation and bare soil the corresponding percentages are 20% and 35% (Miralles et al., 2011; Santanello and Friedl, 2003; Kustas a . Altogether, the fraction of net radiation assumed to be converted into the ground heat flux is the weighted average of the former percentages with the fractional land covers.

2.4 Balloon soundings

The Integrated Global Radiosonde Archive (IGRA; Durre et al. (2006)) is a data set of direct atmospheric sounding observations from balloons across the globe, representative of different environmental and climate conditions (Wouters et al., 2019) and can be used to evaluate ~~the turbulent fluxes from~~ estimated profiles of atmospheric properties. The data set will be used here to evaluate atmosheric profiles derived from forcing an ABL model (Section 2.6) with ERA5 ~~and ERA-I at the catchment scale and yearly temporal resolution for the period 1983–2014.~~ data. The balloon soundings are screened for the observation time and quality as in Wouters et al. (2019). A detailed description of this data set, together with a description of the processing and quality checks can be found in Wouters et al. (2019). The data set used in this study consists of approximately 18000 quality-checked morning–afternoon sounding pairs from 121 locations across the globe from 1981 to 2018.

2.5 GLEAM

The Global Land Evaporation Amsterdam Model (GLEAM) is a process-based semi-empirical model designed to estimate terrestrial evaporation and its separate components at the global scale from satellite observations alone (Miralles et al., 2011). In summary, GLEAM first calculates potential evaporation using the Priestley and Taylor equation (Priestley and Taylor, 1972) for four land cover fractions per grid cell: (1) low vegetation, (2) tall vegetation, (3) bare soil, and (4) open water. Estimates of potential transpiration (for the first two fractions) are converted into actual transpiration by applying an empirical multiplicative stress factor. The latter is calculated as a function of vegetation optical depth – which is used as a proxy for vegetation water content (Liu et al., 2013, 2011) – and root-zone soil moisture. The root-zone soil moisture in GLEAM is calculated using a multi-layer soil water balance model driven by precipitation, and is further optimised using a Newtonian Nudging data assimilation scheme (Martens et al., 2017, 2016). For the bare soil fraction, the evaporative stress factor is calculated based on surface soil moisture alone, while for the open-water fraction, no evaporative stress is considered (i.e. actual equals potential evaporation). Finally, for grid cells covered by snow, sublimation is calculated using the Priestley and Taylor equation with a specific set of parameters (Murphy and Koop, 2005). The fraction of precipitation intercepted by the vegetated surface and directly evaporated back into the atmosphere (i.e. rainfall interception loss) is only calculated for the fraction of tall vegetation~~in GLEAM~~. For this purpose, the implementation of Gash's analytical model of rainfall interception~~loss~~ (Gash, 1979) by Valente et al. (1997) is used~~in GLEAM~~. Ultimately, the total evaporative flux is ~~then~~ calculated by summing the fluxes calculated for the four cover fractions. For a detailed description of GLEAM, we refer the readers to Martens et al. (2017, 2016) and Miralles et al. (2011, 2010).

Here, GLEAM is used as a tool to assess quality differences in some key meteorological drivers of the turbulent fluxes, derived from ERA5 and ERA-I, and to explore the skill of the land-surface model implemented in ERA5 (H-TESSEL) to accurately model the control of the land surface on the turbulent heat fluxes. To do so, GLEAM is forced by an up-to-date version of the GLEAM v3a forcing data base described in Martens et al. (2017), which uses near-surface air temperature and surface net radiation from ERA-I (hereafter referred to as GLEAM+ERA-I). Next, GLEAM is also forced using the same data set, but with near-surface air temperature and surface net radiation from ERA5 (hereafter referred to as GLEAM+ERA5). Although GLEAM has been designed to target the accurate estimation of terrestrial evaporation (or surface latent heat flux)~~only, here, the~~ , we also calculate the estimated surface sensible heat flux ~~, calculated~~ as the residual of the energy balance, ignoring

changes in energy storage (Eq. 2)~~, and~~ . Based on the estimates of both turbulent fluxes, the Bowen ratio from GLEAM ~~are also evaluated. GLEAM~~ is also calculated. The model is run for the period 1989–2015 – where 1989 is used as a spin-up year (Martens et al., 2017) – ~~and the output from both experiments is evaluated against the eddy-covariance data described in Sect. 2.2 at the daily temporal scale, as GLEAM does not support sub-daily simulations~~at daily temporal resolution, and on a regular 0.25° latitude–longitude grid (Martens et al., 2017). All inputs, either sourced from ERA-I or ERA5, are processed as in Martens et al. (2017), including a linear re-sampling in both time and space to the spatio-temporal resolution used by GLEAM.

2.6 CLASS4GL~~boundary layer model~~

The Chemistry Land-surface Atmosphere Soil Slab (CLASS) model for GLobal studies (CLASS4GL; http://class4gl.eu) is a free software tool designed to investigate the dynamics of the ~~atmospheric boundary layer (ABL)~~ ABL and its sensitivity to different land and atmospheric conditions using data from weather balloons (Wouters et al., 2019). The core of ~~the platform~~ CLASS4GL is the ABL model CLASS, which is coupled to a soil-vegetation module ~~that simulates the~~ allowing the simulation of the diurnal evolution of the ABL ~~using a timestep~~ with a temporal resolution of 60 seconds. The platform is able to mine appropriate observational data from global radio soundings, satellite data, and reanalysis data from the last 40 years to constrain and initialise the ABL model. Its interactive interface automatises multiple simulations of the ~~atmospheric boundary layer~~ ABL in parallel and allows to perform global perturbation experiments. It aims ~~at fostering~~ to foster a better understanding of land–atmosphere feedbacks and to disentangle the drivers of (extreme) weather conditions globally.

Here, CLASS4GL is used as a tool to assess whether the surface energy partitioning in ERA5 has been improved upon ~~the one in~~ ERA-I ~~.~~ in a similar experiment as described in Sect. 2.5 with GLEAM. Therefore, CLASS4GL is forced with the turbulent fluxes derived from both ERA5 and ERA-I to simulate diurnal tendencies of potential temperature, humidity, and mixed-layer height. ~~The latter are evaluated against direct observations from balloon soundings from across the globe, representative of different environmental and climate conditions (Wouters et al., 2019). These soundings are sourced from the Integrated Global Radiosonde Archive (IGRA; Durre et al. (2006)) and are screened for the observation time and quality.A detailed description of this~~ As described by Wouters et al. (2019), the evaporative fraction derived from reanalysis data (either ERA-I or ERA5) is used to guide the simulations of the ABL diurnal evolution, and the resulting afternoon profiles of humidity, potential temperature, and ABL height.

2.7 Evaluation strategy

2.7.1 Evaluation using eddy-covariance data and balloon soundings

Both the turbulent fluxes (and Bowen ratio) from the reanalyses (Sect. 2.1) and the estimates from the GLEAM experiments (Sect. 2.5) are directly compared against the in situ eddy-covariance measurements (Sect. 2.2). For each eddy-covariance site in the validation data set, ~~together with a description of the processing and quality checks can be found in Wouters et al. (2019) .The data set used in this study consists of 18000 quality-checked morning–afternoon sounding pairs from 121 locations across~~

~~the globe from 1981 to 2018.~~ the variables from the overlapping model grid cells are extracted at their native spatial resolution and both as 3-hourly and daily (temporal) aggregates. Note that for the experiments involving GLEAM, only daily estimates are available. Eddy-covariance sites located within the same model grid cell are treated separately in the validation to avoid potential problems resulting from merging sensors with different absolute values and gaps in their record (Martens et al., 2017) . Also note that there is a substantial mismatch between the footprint of the eddy-covariance system and the model grid cells, resulting in a representativeness error that can be a substantial fraction of the total error (Jiménez et al., 2018).

~~In summary, the evaporative fraction from both ERA-I and ERA5 is used to optimise the root-zone soil moisture in CLASS4GL in an iterative procedure as described in Wouters et al. (2019). The resulting output of the modelling framework is evaluated against the measurements~~ As the temporal variability of the turbulent fluxes is strongly influenced by the seasonal cycle of its main drivers at the scales considered in this experiment, the performance of the land-surface schemes in response to anomalous weather conditions (i.e. with respect to the seasonal cycle) might be masked when raw time series are analysed. As such, the evaluation of the turbulent fluxes against the FLUXNET data set will be done based on standardised anomalies to better evaluate the skill of the reanalyses in capturing the effect of specific meteorological conditions on the surface energy partitioning. Therefore, standardised anomalies of the turbulent fluxes are calculated (and Bowen ratio) from (1) the reanalyses, (2) the GLEAM experiments, and (3) the eddy-covariance measurements prior to calculating validation metrics. Note that the calculation of standardised anomalies allows to directly compare the quality of the turbulent fluxes and the Bowen ratio, despite their different orders of magnitude.

Anomaly time series are calculated by (1) subtracting for each time interval the expected value (i.e. the climatology), calculated as the multi-annual average for that time interval, and (2) dividing by the standard deviation of the expectation. To calculate climatologies of the eddy-covariance data, only FLUXNET sites with a minimum record length of five years are considered, resulting in 77 eddy-covariance towers for the evaluation of the anomaly time series (Fig. 1).

Using the standardised anomalies of the in situ eddy-covariance measurements as a reference, the Pearson correlation coefficient (R) and Mean Absolute Difference (MAD) of the reanalyses data sets and the estimates from GLEAM are calculated to evaluate their quality (Sect. 3.1.1). In addition, the Mean Difference (MD) of the raw data series is calculated to assess the bias in the estimates. Metrics are visualised in violin plots constructed using a kernel density estimation approach with a band width calculated according to Scott (1979). For the MD and R, a 95% confidence interval is calculated at each FLUXNET site following the procedure outlined in De Lannoy and Reichle (2016). First, the temporal auto-correlation in both the reference and estimated time series is calculated to correct the degrees of freedom (Gruber et al., 2020). Second, a confidence interval is calculated at each FLXNET site assuming a normal distribution for R (after applying a Fisher Z-transformation to the time series) and a Student t-distribution for the MD. Metrics are then assumed to be statistically different at the 5% significance level if their confidence intervals do not overlap. Note that we do not calculate confidence intervals for the MAD, as there are no analytical solutions available for this metric and the calculation thus requires a non-parametric approach relying on computationally heavy Monte Carlo simulations (Gruber et al., 2020). Finally, the confidence intervals for the MD and R are averaged across the FLUXNET data set and the average confidence interval is reported.

In a similar manner as for the GLEAM experiment, the simulations of CLASS4GL (Sect. 2.6) are validated against afternoon profiles from balloon soundings ~~thereby providing an indirect and independent~~ sourced from the IGRA data set (Sect. 2.4). However, the skill of CLASS4GL is evaluated based on the Root Mean Squared Error (RMSE) – rather than MAD – R, and MD, all calculated on raw time series, and results are visualised in Taylor plots.

2.7.2 Evaluation using catchment energy-balance data

Next to the evaluation of the ~~surface energy partitioning in both reanalyses . Alltogether, the experiment investigates whether the partitioning provided by~~ turbulent fluxes from ERA5 ~~is more consistent with observed atmospheric boundary layer parameters than that provided by ERA-I~~against in situ eddy-covariance measurements, an evaluation against catchment-scale water and energy balance data (Sect. 2.3) is also performed. Given the typical bias in eddy-covariance measurements, especially in case of the surface latent heat flux (Beer et al., 2010), an evaluation of the magnitude of the fluxes should be interpreted with care. On the other hand, the catchment-scale energy-balance data is thought to be less biased, especially at the temporal scales considered in this study, and is therefore better suited to evaluate the magnitude of the fluxes (Miralles et al., 2016).

For each catchment in the data set, the turbulent fluxes of the reanalyses (Section 2.1) are temporally aggregated to the annual resolution and spatially aggregated to the scale of the catchments. Next, the MD between the reference data set and the reanalysis is calculated to assess the magnitude of the surface energy partitioning. Results are spatially visualised in global maps and compared against each other by means of scatter plots.

3 Results and discussion

3.1 Evaluation using eddy-covariance data

3.1.1 Direct comparison to in situ data

Figure 2 shows violin plots of the ~~Mean Difference (MD ,~~ MD (raw in situ time series as reference), ~~Mean Absolute Difference (MAD ,~~ MAD (anomaly in situ time series as reference), and ~~Pearson correlation coefficient (~~R ~~,~~(anomaly in situ time series as reference) of the turbulent fluxes and the Bowen ratio against in situ eddy-covariance measurements. ~~Violins are shown~~ Average metrics across the FLUXNET data set and their confidence interval are reported in Table 1. Violin plots are presented for the surface latent heat flux (3-hourly and daily resolution), surface sensible heat flux (3-hourly and daily resolution) and Bowen ratio (daily resolution), for ERA5 (green~~; GLEAM+ERA5~~) and ERA-I (~~orange; GLEAM+ERA-I~~yellow), respectively. As shown, statistics are consistently (and statistically significantly) better for ERA5 than for ERA-I, with typically higher R and lower MAD against in situ measurements~~(the bias~~ , even though the bias (MD) remains relatively similar~~)~~. This indicates that ERA5 is better capturing the temporal dynamics in surface energy partitioning, both at sub-daily and daily temporal resolutions. Especially for the daily-aggregated surface sensible heat flux, a clear improvement ~~is shown~~can be seen, with the median R of ERA5 across all reference sites approaching the 75% percentile of the ERA-I distribution. Nevertheless, differences are statistically significant in more sites at the sub-daily scale than at daily resolutions: the Pearson correlation coefficient for the

surface sensible heat flux from ERA5 is significantly better (at the 5% significance level) at 63% and 38% of the sites at the 3-hourly and daily temporal resolutions, respectively. ERA-I, on the other hand, is only significantly better in approximately 10% of the sites, while in the remainder of sites, differences are not significant. For the surface latent heat flux and Bowen ratio, improvements are less remarkable, but still consistent~~.~~, as R is significantly better for ERA5 in 59%, 29%, and 39% of the eddy-covariance sites for the surface latent heat flux (3-hourly and daily resolutions) and the Bowen ratio. The opposite is only true in about 8% of the sites. As shown in Fig. 2, both ERA5 and ERA-I tend to overestimate the surface latent heat flux and underestimate the Bowen ratio. Conversely, the average bias in the surface sensible heat flux is close to zero. However, advances in ERA5 have not been able to make a huge difference in these tendencies, as statistics of ERA5 and ERA-I are close to each other and statistically significant in only 1–2 sites. Notably, for both ERA-I and ERA5, validation statistics are generally better for sensible than for latent heat fluxes (see higher median R and lower MAD for sensible heat fluxes, irrespective of data set and temporal aggregation). Albeit the differences in pre-processing techniques and in the sample of eddy-covariance sites, these results are consistent with those by Balsamo et al. (2015) based on a validation of ERA-I only. When the seasonality is not removed (Fig. A1), turbulent fluxes of ERA5 still outperform those from ERA-I, although differences are smaller. In terms of seasonal cycle, the surface sensible heat flux is not necessarily better estimated than the surface latent heat flux; in fact, statistics are generally worse at daily temporal resolution as shown in Fig. A1.

~~Violin plots of temporal validation statistics of the surface latent heat flux ($\lambda\rho E$), surface sensible heat flux (H), and Bowen ratio (β) from ERA5 (green) and ERA-I (orange). Statistics are calculated against in situ eddy-covariance measurements at both 3-hourly and daily temporal resolutions. Violins represent the distribution of the individual validation statistics with indication of the median and inter-quartile range and are calculated using a kernel density estimation approach. Statistics include the Mean Difference (MD, raw in situ time series from 143 sites as reference), Mean Absolute Difference (MAD, anomaly in situ time series from 77 sites as reference), and the Pearson correlation coefficient (R, anomaly in situ time series from 77 sites as reference). For MD, the distribution of β is plotted on the right y-axis.~~

Figure 3 shows the difference between temporal validation statistics calculated at the anomaly time series (i.e. MAD and R) of the surface latent heat flux, surface sensible heat flux, and Bowen ratio from ERA5 and ERA-I. Sites are clustered as a function of ~~precipitation and mean~~ mean annual precipitation and near-surface air temperature measured at the corresponding eddy-covariance ~~station~~site. Results are consistent with those in Fig. 2, with an overall higher quality (~~blue~~ green color) in the sensible and latent heat fluxes from ERA5. However, it can be argued that there is a tendency of ERA-I to perform better than ERA5 in warm and dry regimes, especially for the latent heat flux and Bowen ratio. These climates are, nonetheless, only sampled by three eddy-covariance towers and thus results may not be generalised. In addition, conclusions based on the performance in certain climate regimes should be interpreted with care, as FLUXNET sites are not uniformly distributed~~,~~: mild climates are generally over-represented ~~,~~and most sites are located in Europe and the ~~Continental United States (CONUS)~~ CONUS, as shown in Fig. 1 and described in Baldocchi et al. (2001).

~~Difference between temporal validation statistics of the surface latent heat flux ($\lambda\rho E$), surface sensible heat flux (H), and Bowen ratio (β) from ERA5 and ERA-I grouped as a function of precipitation rate (P) and near-surface air temperature (T) calculated at the in situ site. Statistics are calculated against in situ eddy-covariance measurements at daily resolution and~~

then averaged across the sites within each group. Statistics include the Mean Absolute Difference (MAD, anomaly in situ time series from 77 sites as reference) and the Pearson correlation coefficient (R, anomaly in situ time series from 77 sites as reference). Circles show the R from ERA5 minus the one from ERA-I, while squares show the MAD from ERA-I minus the one from ERA5; hence, blue colors represent better statistics for ERA5 compared to ERA-I. The size of the symbols relates to the number of in situ sites per group.

Presumably, much of the improvement in surface energy partitioning in ERA5 over ERA-I can be attributed to a better representation of land-surface processes in the more advanced H-TESSEL land-surface model . and the improved data assimilation system wrapped around the model. Note that both improvements in the atmospheric data assimilation system (by improving the atmospheric drivers of the turbulent fluxes) and the land-surface data assimilation (by improving the land-surface constraint on the turbulent fluxes) might affect the turbulent fluxes. The better performance of H-TESSEL – in reference to TESSEL, its antecessor used in ERA-I – was already illustrated by Balsamo et al. (2015), who compared the quality of different land-surface variables from ERA-I and ERA-I/Land over the Northern Hemisphere. The latter ERA-I/Land is in essence an offline simulation of H-TESSEL forced with atmospheric data derived from ERA-I. Although quality differences between ERA-I and ERA-I/Land cannot can not only be attributed to the land-surface scheme but also to the different model set-up (i.e. online vs. offline simulation), Balsamo et al. (2015) argued that most of the improvement was due to the land-surface model. As H-TESSEL is now also implemented in ERA5, similar analogous improvements can thus be expected in ERA5 over ERA-I regarding the simulation of land-surface variables.

Despite the fact that several studies have shown the high performance of H-TESSEL as compared to TESSEL for simulating a variety of land-surface parameters (e.g. Balsamo et al., 2015; Albergel et al., 2012), Balsamo et al. (2015) also showed that improvements in the turbulent fluxes of ERA-I/Land over ERA-I could not be uniquely linked to the different land-surface scheme. Hence, the better quality of surface energy partitioning in ERA5 does most likelynot only benefit from is, most likely, not only owed to an improved parameterisation of the land surface, but also from a better quality of the atmospheric drivers, simulated by the coupled atmospheric model, which is constrained by a 4D-variational data assimilation algorithm benefiting from a larger number of quality controlled observations (Hersbach et al., 2020, 2018)of a large number of quality-controlled observations (Hersbach et al., 2020). The better quality of some key meteorological parameters is confirmed by the results presented in Fig. A3, which shows violin plots of the validation statistics for surface net radiation, 2-meter air temperature, and precipitation at the FLUXNET sites, for 3-hourly and daily temporal resolutions, respectively. Although statistics from ERA5 are better at both temporal resolutions, especially the sub-daily variability of all three variables has been substantially improved over ERA-I, which may largely be the result of a better modelling of cloud properties in ERA5 (Hersbach et al., 2020, 2018) (Hersbach et al., 2020).

Finally, as described in Sect. 2.1, one of the key improvements in ERA5 upon its predecessor is the higher spatial resolution at which atmospheric and land processes are resolved. However, Fig. A2 shows that when ERA5 is linearly re-sampled to the spatial resolution of ERA-I, statistics calculated against eddy-covariance measurements only change marginally. Nevertheless, such an analysis only gives a crude idea of the impact of the spatial resolution as (1) due to non-linear processes and feedback mechanisms, a simple re-sampling of the model output does not properly represent the effect of the high-resolution

numerical modelling~~;~~ , (2) the effect is expected to be the highest in complex terrain such as mountainous regions, coastal areas, or highly-heterogeneous landscapes, which are under-represented in the FLUXNET data base~~;~~ , and (3) ~~representativity~~ representativeness errors – resulting from the relatively small footprint of eddy-covariance towers as compared to model grid cells – ~~are still~~ remain considerable at the spatial resolution of ERA5.

3.1.2 Evaluation using GLEAM

Forcing GLEAM with meteorological data derived from ERA5 and ERA-I provides a convenient and alternative means to evaluate and compare the quality of the reanalyses. Moreover, it allows an evaluation of the usefulness of ERA5 to drive offline models explicitly designed to estimate ~~specific~~ land-surface ~~variables (in the~~ fluxes (in case of GLEAM, terrestrial evaporation). Nevertheless, results of such an experiment should be interpreted with care as errors in the forcing might be compensated for by the model. However, parameters in GLEAM are fully based on literature studies (Martens et al., 2017; Miralles et al., 2011) and are not calibrated, the analysis presented in this study is performed over a large number of sites, and the modelling concepts of GLEAM and ERA-I/ERA5 are substantially different. Hence, it is assumed here that errors in neither ERA-I, nor ERA5 are compensated for by GLEAM.

Figure 4a shows violin plots of the MD (raw in situ time series as reference), MAD (anomaly in situ time series as reference), and R (anomaly in situ time series as reference) of the turbulent fluxes and the Bowen ratio derived from GLEAM against in situ eddy-covariance measurements. ~~Violins are~~ The average R and MD, together with their confidence interval, are reported in Table 1. Violin plots are shown for both turbulent fluxes and the Bowen ratio at daily temporal resolution; the violin limbs correspond to GLEAM forced with ERA5 (green) and ERA-I (~~orange~~yellow), respectively. Results presented in Fig. 4a show that the estimates of the surface latent heat flux from GLEAM+ERA5 are consistently better than those from GLEAM+ERA-I, especially in terms of R and MAD, while the bias in both is comparable and close to zero on average. While for the MD, GLEAM+ERA5 is only significantly better in a handful of sites, R is significantly better in 22% of the sites for the turbulent fluxes, and in 3% of the sites for the Bowen ratio. However, in the majority of sites (75% for the turbulent fluxes and 91% for the Bowen ratio), differences in R are not statistically significant. These findings support the ones discussed in Sect. 3.1.1, where it was found that some key meteorological drivers of the surface turbulent fluxes are in fact better represented in ERA5 than in ERA-I. On the other hand, with the exception of the bias, statistics for the surface sensible heat flux and Bowen ratio are slightly worse for GLEAM+ERA5 than for GLEAM+ERA-I, but not statistically significant in terms of R, as evidenced by the percentages reported above. Nonetheless, when the seasonal cycle is not removed prior to the analysis (Fig. A4a) GLEAM+ERA5 performs consistently (albeit only slightly) better for all variables, suggesting that the seasonality of the meteorological variables used to force GLEAM is better captured in ERA5 than in ERA-I. Despite the fact that the most prominent differences in quality of the surface latent heat flux from GLEAM+ERA5 and GLEAM+ERA-I can be found in mild climates as indicated in Fig. 5a, there is no clear tendency of GLEAM+ERA5 to perform better under specific climatic conditions. The surface sensible heat flux and Bowen ratio from GLEAM+ERA5, on the other hand, tend to degrade in quality (compared to GLEAM+ERA-I) when the climate gets drier and colder. It should be emphasised here again that GLEAM has been specifically designed to estimate the latent heat flux, thus the surface sensible heat flux – calculated here as the residual

from the energy balance – has not been subject to equally extensive validations than its latent counterpart, and is prone to be more uncertain.

Violin plots of temporal validation statistics of the surface latent heat flux ($\lambda\rho E$), surface sensible heat flux (H), and Bowen ratio (β) from GLEAM+ERA5, GLEAM+ERA-I, and ERA5. **(a)** Compares the violins from GLEAM+ERA5 and GLEAM+ERA-I and **(b)** compares the violins from GLEAM+ERA5 and ERA5. Statistics are calculated against in situ eddy-covariance measurements at daily temporal resolution. Violins represent the distribution of the individual validation statistics with indication of the median and inter-quartile range and are calculated using a kernel density estimation approach. Statistics include the Mean Difference (MD, raw in situ time series from 143 sites as reference), Mean Absolute Difference (MAD, anomaly in situ time series from 77 sites as reference), and the Pearson correlation coefficient (R, anomaly in situ time series from 77 sites as reference). For MD, the distribution of β is plotted on the right y-axis.

Difference between temporal validation statistics of the surface latent heat flux ($\lambda\rho E$), surface sensible heat flux (H), and Bowen ratio (β) from GLEAM+ERA5, GLEAM+ERA-I, and ERA5 grouped as a function of precipitation rate (P) and near-surface air temperature (T) calculated at the in situ site. **(a)** Compares the statistics from GLEAM+ERA5 and GLEAM+ERA-I and **(b)** compares the statistics from GLEAM+ERA5 and ERA5. Statistics are calculated against in situ eddy-covariance measurements at daily resolution and then averaged across the sites within each group. Statistics include the Mean Absolute Difference (MAD, anomaly in situ time series from 77 sites as reference) and the Pearson correlation coefficient (R, anomaly in situ time series from 77 sites as reference). In **(a)** circles show the R from GLEAM+ERA5 minus the one from GLEAM+ERA-I, while squares show the MAD from GLEAM+ERA-I minus the one from GLEAM+ERA5; hence, blue colors represent better statistics for GLEAM+ERA5 compared to GLEAM+ERA-I. In **(b)**, statistics from GLEAM+ERA-I are replaced by ERA5; hence, blue colors represent better statistics for GLEAM+ERA5 compared to ERA5. The size of the symbols relates to the number of in situ sites per group.

The turbulent fluxes and Bowen ratio from GLEAM+ERA5 can also be directly compared to ERA5 to provide a crude evaluation of the skill of H-TESSEL as compared to the simpler land-surface scheme in GLEAM. Figure 4b shows that ERA5 is better capturing the temporal dynamics of the anomalies, generally resulting in lower MAD and higher R for all variables. In terms of R, ERA5 is performing significantly better (at the 5% significance level) at 27%, 39%, and 27% of the sites for the surface latent heat flux, surface sensible heat flux, and Bowen ratio, respectively. GLEAM+ERA5 is only performing better in 15%, 9%, and 18% of the sites for the same variables, while in the majority of sites, differences are not significant. Only in terms of the bias, ERA5 is overall performing worse than GLEAM+ERA5 ~~,~~ (but again, only significant at a very limited number of sites), especially for the surface latent heat flux, which is consistently overestimated in ERA5 for almost all in situ sites (close to 75% of the sites have a positive bias, Fig. 4b). This results in a median MD of 9 W m^{-2} compared to the slight underestimation of -2 W m^{-2} for GLEAM+ERA5 at daily time scales. The positive bias in the surface latent heat flux from ERA5 is very similar to the one from ERA-I, with a median MD of 10 W m^{-2} across all in situ sites at daily resolutions (Fig. 2). The tendency to overestimate the latent heat flux in ERA-I has been previously reported in different studies (Michel et al., 2016; Miralles et al., 2016; Balsamo et al., 2015; Decker et al., 2012), and important changes in the IFS have thus not been able to mitigate this bias in ERA5. Given the interaction between the coupled atmospheric and land-surface model in the

reanalysis, the consistent positive bias in the surface latent heat flux is potentially affected by both components of the modelling framework. Although it is hard to identify the exact cause of this bias, it might be induced by the overestimation of the number of wet days typically found in reanalysis data sets (Beck et al., 2019), combined with precipitation rates that are often underestimated (Beck et al., 2019), and vegetation density that might be overestimated (Král, 2011). This presumably results in an overestimation of the interception loss (Král, 2011), an important component of the total latent heat flux in densely-vegetated regions (Martens et al., 2017; Miralles et al., 2010). ~~This hypothesis is somehow~~ Note that this hypothesis is partially supported by our analysis: despite the fact that a positive bias can be found virtually everywhere, the strongest biases are typically found in densely-vegetated sites (not shown). We should ~~emphasis~~ emphasise here, however, that biases calculated against eddy-covariance measurements ~~should~~ have to be interpreted with care, given ~~the representativity errors~~ representativeness errors resulting from the mismatch in spatial footprint between the grid cell and the instrument, and provided that turbulent heat fluxes are thought to be generally underestimated by the eddy-covariance technique, especially in case of the surface latent heat flux (Beer et al., 2010). When the seasonal cycle is not removed prior to the evaluation (Fig. A4b), GLEAM+ERA5 seems to perform equally good or slightly better than ERA5, indicating that GLEAM+ERA5 is marginally better than ERA5 at capturing the seasonal dynamics (Fig. A4b), but worse at capturing the response of surface energy partitioning to short-term anomalies in meteorological conditions (Fig. 4b). Nevertheless, we would like to highlight that ERA5 is a fully-coupled land-atmosphere system permitting a feedback from the land surface towards the atmosphere, while GLEAM is an offline land-surface model forced with atmospheric variables from ERA5. We note that this coupling between the land surface and the atmosphere might have a substantial impact on the quality of the turbulent fluxes (Draper et al., 2018; Balsamo et al., 2015), potentially explaining the differences between GLEAM+ERA5 and ERA5.

~~Figure~~Nonetheless, Fig. 5b shows that for the surface latent heat flux, the better performance of ERA5 over GLEAM+ERA5 is mainly due to its better statistics in relatively wet or cold climatic regimes. In drier regimes and, especially warm regions (mainly located along the west coast of the CONUS and few eddy-covariance sites in Australia; Fig. 1), GLEAM+ERA5 seems to better capture the anomalies of the surface latent heat flux, which might indicate that H-TESSEL has room to improve the response to water stress. For the Bowen ratio, similar conclusions may be drawn, even though the quality of the sensible heat flux in ERA5 is consistently better than in GLEAM+ERA5.

3.2 Evaluation using catchment energy-balance data

As described in Sect. 2.3, observations of river discharge may be combined with precipitation, net radiation, and ~~the~~ ground heat flux to derive catchment-scale and long-term estimates of the surface turbulent fluxes and the Bowen ratio, providing an alternative means to evaluate the surface energy partitioning in ERA-I and ERA5. Figure 6 compares the percentage MD (%MD, i.e. MD divided by the mean of the reference data set) of the surface latent heat flux, surface sensible heat flux, and Bowen ratio (observations of catchment-scale variables as reference) from ERA5 ~~to the one from~~ and ERA-I ~~by means of~~ using a scatter plot. The results shown in Fig. 6 largely correspond to the ones shown in Fig. 2 for the MD and point again to a substantial overestimation of the surface latent heat flux from ERA-I; in 83% of the catchments, a positive bias is obtained~~for the flux. The~~ . Conversely, the surface sensible heat flux ~~on the other hand~~ is generally underestimated (a

negative bias is found in 61% of the catchments), resulting in an underestimation of the catchment-scale Bowen ratio as well (a negative bias is found in 80% of the catchments). While absolute biases for the surface latent heat flux from ERA5 are lower than from ERA-I (an improvement is found in 75% of the catchments), ERA5 still overestimates the flux in most catchments~~. Strikingly, ~~, as also indicated by Hersbach et al. (2020). More striking are the results for the surface sensible heat flux: while ERA-I generally underestimates the ~~surface sensible heat~~ flux, ERA5 overestimates ~~the flux~~ it in about 70% of the catchments. In addition, the absolute bias of the surface sensible heat flux from ERA5 is higher than in ERA-I in 55% of the catchments. However, this potential overestimation is not confirmed by the in situ validation presented in Sect. 3.1.1 (Fig. 2), where the surface sensible heat flux from both reanalyses appeared nearly unbiased. Finally, for the Bowen ratio, estimates of ERA5 ~~are~~ appear better in about 60% of the catchments, arguably reflecting the improvement in the surface latent heat flux. Note that a rather strong overestimation of the surface latent heat flux was also found in other reanalyses such as NASA's MERRA and MERRA-2 (Draper et al., 2018). However, in the latter reanalyses, both surface turbulent fluxes were consistently overestimated which could potentially be linked to a positive bias in the incoming radiation at the land surface.

~~Scatter plot of the bias of the surface latent heat flux ($\lambda\rho E$), surface sensible heat flux (H), and Bowen ratio (β) from ERA-I versus ERA5. The bias is calculated against catchment-scale estimates of the fluxes derived using discharge data (Eqs. 1–3) and is assessed by the percentage Mean Difference (%MD, raw time series from 707 catchments as reference). The green area indicates points where the bias in ERA5 is better than in ERA-I, and vice versa for the orange area.~~

Figures 7–9 show maps of the catchment-scale %MD of the surface latent heat flux (Fig. 7), surface sensible heat flux (Fig. 8), and Bowen ratio (Fig. 9) for ERA5, ERA-I, and the difference in their absolute values. While ERA-I overestimates the surface latent heat flux virtually everywhere, biases are relatively larger in the east of the CONUS and the south of Europe (in regions like Spain and the south of France). In these regions, a strong reduction in bias can be observed for ERA5. Despite the complex interactions between the land surface and the atmosphere in the IFS, these improvements can potentially be related to an improved representation of precipitation in ERA5 as shown by Hersbach et al. (2020) and affecting (1) interception loss in radiation-limited regions such as the east of the CONUS – which might represent a substantial portion of total evaporation in forested regions (Martens et al., 2017; Miralles et al., 2011, 2010) – and (2) the land-surface constraint on terrestrial evaporation in water-limited evaporation regimes like the south of Europe. Note that the latent heat flux in the latter regions will also be strongly affected by improvements in the land-surface data assimilation system (Hersbach et al., 2020; Balsamo et al., 2 . Over large parts of Europe and western Russia on the other hand, the surface latent heat flux from ERA5 is nearly unbiased, while the overestimation in other regions still remains, albeit reduced compared to ERA-I. Except for a small number of catchments in the northeast of Brazil and the west of the Sahel, the bias of the surface latent heat flux is lower in ERA5 than in ERA-I. The surface sensible heat flux from ERA-I is typically underestimated in high latitudes and the eastern part of the CONUS, while an overestimation can be seen in most other regions. However, as discussed in the previous paragraph, the bias in the surface sensible heat flux of ERA5 is typically higher, especially over Europe, western Russia, and the east of the CONUS, regions where the bias in the surface latent heat flux ~~improved.~~ is reduced in ERA5. Finally, in absolute terms, the bias in the Bowen ratio increases from ERA5 to ERA-I as evidenced in Fig. 9, and largely follows the patterns set by the bias in the surface sensible heat flux (Fig. 8).

~~Maps of the bias of the surface latent heat flux ($\lambda\rho E$) from ERA5 and ERA-I. The bias is calculated against catchment-scale estimates of the fluxes derived using discharge data (Eqs. 1–3) and is assessed by the percentage Mean Difference (%MD, raw time series from 707 catchments as reference). In the last row, Δ presents the difference between the absolute bias in ERA5 and ERA-I; hence, blue colors represent lower bias in ERA5 than in ERA-I.~~

~~Maps of the bias of the surface sensible heat flux (H) from ERA5 and ERA-I. The bias is calculated against catchment-scale estimates of the fluxes derived using discharge data (Eqs. 1–3) and is assessed by the percentage Mean Difference (%MD, raw time series from 707 catchments as reference). In the last row, Δ presents the difference between the absolute bias in ERA5 and ERA-I; hence, blue colors represent lower bias in ERA5 than in ERA-I.~~

~~Maps of the bias of the Bowen ratio (β) from ERA5 and ERA-I. The bias is calculated against catchment-scale estimates of the fluxes derived using discharge data (Eqs. 1–3) and is assessed by the percentage Mean Difference (%MD, raw time series from 707 catchments as reference). In the last row, Δ presents the difference between the absolute bias in ERA5 and ERA-I; hence, blue colors represent lower bias in ERA5 than in ERA-I.~~

Finally, it should be emphasised here that the quality of the catchment-scale sensible heat flux (and Bowen ratio) estimates used as reference is potentially lower than that of the surface latent heat flux, as (1) the assumption that the ground heat flux is a fixed fraction of the surface net radiation only affects the estimates of the sensible heat (Eq. 2), and (2) the estimates of sensible heat flux depend on the estimates of surface latent heat (Eq. 2), resulting in a propagation of errors which is difficult to assess. Hence, the catchment-scale evaluation of the surface sensible heat flux ~~should be~~ and Bowen ratio should be more carefully interpreted.

3.3 Evaluation using CLASS4GL

Figure 10 shows the ~~evaluation of the model output~~ validation of the estimated afternoon ABL properties from CLASS4GL ~~– forced with~~ forced with the surface energy partitioning from ERA-I (on the one hand) and ERA5 (on the other hand~~– against the~~). The validation is performed by comparison against a global archive of balloon soundings (Sect. 2.6). Results are shown for the diurnal ~~tendency~~ temporal change (tendency) of potential temperature ($d\theta/dt$), humidity (dq/dt), and mixed-layer height (dh/dt). ~~It is found that the overall model performance for~~ The overall performance at reproducing the diurnal ABL tendencies is improved when CLASS4GL is forced with ERA5 instead of ERA-I. This is the case for all ~~scores~~ statistical scores being considered and for each ~~variable~~ ABL variable being analysed. In addition, this is also the case ~~for most of the~~ in most Köppen–Geiger climate classes~~separately~~, which suggests that the higher performance is consistent ~~among the regions worldwide~~across climate regimes. The largest ~~impact is found on the simulated~~ improvement in simulated ABL properties is found for the tendency of specific humidity, where the bias is reduced from ~~0.12 to 0.06~~0.10 to 0.05 g kg^{-1} h^{-1} ~~if~~ when CLASS4GL is forced with ERA5 instead of ERA-I~~, respectively. Most~~ . Most of the improvement can be found in days where the mixed layer tends to dry out during ~~its daytime~~ the diurnal growth (i.e. negative tendency of specific humidity) and is most likely related to the substantially lower bias in ~~the~~ surface latent heat flux from ERA5~~than from ERA-I~~ , as discussed in Sects. 3.1.1 and 3.2. Also, the Pearson correlation coefficient (~~0.38 vs. 0.52~~0.37 vs. 0.50), normalised Root Mean Squared Error (~~RMSD~~RMSE; 0.22 vs. ~~0.16~~0.17 g kg^{-1} h^{-1}), and normalised standard deviation (1.2 vs. 1.03) point towards ~~improvement~~

~~of the model performance~~improvements of the ABL simulations when forced by the surface energy partitioning from ERA5. For the other variables (dh/dt and $d\theta/dt$)~~improvements are less clear, but~~ , improvements are only minor, but still consistent. These results highlight that the surface energy partitioning in ERA5 ~~improves the skill of CLASS4GL in simulating the diurnal variations in the atmospheric boundary layer; hence beneficial for boundary layer climate studies~~can lead to improved skill in the diurnal ABL simulations by mixed-layer models such as CLASS.

~~Skill comparison of the CLASS4GL model (http://class4gl.eu; Wouters et al., 2019) for reproducing diurnal changes in boundary layer properties forced with surface evaporative fraction from ERA5 versus ERA-I. Shown are the tendencies of the mixed-layer height (dh/dt), potential temperature ($d\theta/dt$) and specific humidity (dq/dt), which are assessed by comparison of model simulations against the IGRA sounding data between 1981 and 2015. The first (ERA5 forced) and second (ERA-I forced) row show modeled versus observed data points (gray) and the corresponding median (green) and interquartile range (red) of the model. The 1–1 line is shown as a black dashed line. The last row indicates the model skill forced with ERA5 (circles) versus ERA-I (squares) with Taylor plots. The transparent symbols show the overall performance of 18k sounding pairs from 121 stations, whereas the colored symbols indicate the performance per Köppen-Geiger climate class and for which the size is proportional to number of sounding pairs.~~

3.4 Global patterns of surface energy partitioning

Figure 11 shows maps of the multi-annual average of the surface latent heat flux, surface sensible heat flux, and Bowen ratio from ERA5 and ERA-I, as well as the difference between both. In both data sets, the expected geographical patterns set by the general climatic conditions emerge. High values for the surface latent heat flux can be found around the ~~Equator~~ equator where both the availability of water and the supply of energy are high, while the lowest values can be found in arid regions such as the Sahara desert, central Australia, the Namibian desert, and the Gobi desert. In terms of surface sensible heat flux, an opposite pattern is shown, with relatively lower values in the tropics, where most of the available energy is consumed to evaporate water, and very high values in the deserts, where virtually no water is evaporated. The Bowen ratio clearly marks the tropical forests and deserts; with intermediate values for mild climates such as central and western Europe.

~~Maps of the multi-annual average of surface latent heat flux ($\lambda\rho E$, W m^{-2}), surface sensible heat flux (H, W m^{-2}), and Bowen ratio (β) from ERA5 and ERA-I. In the last row, Δ presents the difference between ERA5 and ERA-I; hence, blue colors represent higher values in ERA5 compared to ERA-I.~~

The globally-averaged surface sensible heat flux from land amounts to 27.2 W m^{-2} and 26.9 W m^{-2} for ERA5 and ERA-I, respectively; a difference of only 1.1% (ERA-I as reference). For the surface latent heat flux, the difference is higher and sums up to -5.2% (ERA-I as reference), with global averages of 44.1 W m^{-2} and 46.5 W m^{-2} for ERA5 and ERA-I, respectively. The latter two values correspond to a yearly total volume of evaporated water of approximately 97.8·10^3 km^3 and 103.1·10^3 km^3. ~~Despite that these values~~ Similar values typically found in literature – although based on different land-surface models or retrieval algorithms, input data sets, or region considered (e.g. areas permanently covered by snow or ice included or not) – range between 55·10^3 km^3 and 80·10^3 km^3 (Miralles et al., 2016; Wang and Dickinson, 2012, and references therein), pointing towards an overestimation of the total volume of evaporated water in both ERA-I and ERA5. In terms of globally-averaged

energy fluxes, the turbulent fluxes from both reanalyses lie within (or close to) the uncertainty ranges reported by Wild et al. (2015)~~–~~, who inferred the magnitude of the global energy fluxes based on a detailed analysis of a variety of observations and model-based estimates~~– the values for the~~ . However, the surface sensible heat flux ~~and~~ from both reanalyses can be found near the lower boundary of the interval, while the surface latent heat flux ~~are at the lower and higher limit of these intervals, respectively.~~ may be found near the upper limit of the interval. This is also the case when compared to values reported in Draper et al. (2018) who analysed the turbulent fluxes of NASA's reanalyses products MERRA, MERRA2, and MERRA-Land. They found values for both fluxes ranging between 42 W m^{-2} and 50 W m^{-2}, depending on the reanalysis considered. These results confirm our findings in Sects. 3.1 and 3.2 and ~~align~~ are in line with results previously reported in literature (e.g. Miralles et al., 2016; Wild et al., 2015; Mueller et al., 2013; Jiménez et al., 2011, and references therein) where similar biases were found for ERA-I.

Figure 11 shows that the lower globally-averaged surface latent heat flux from ERA5 mainly results from reduced values along the east coast of the CONUS, the south of Europe, the Sahel, India, and large parts of South America. These regions align well with the areas identified in Miralles et al. (2016) where ERA-I seemed to strongly overestimate the surface latent heat flux, and thus point to a better performance of ERA5 in these specific regions, although positive biases still prevail (Fig. 7). The surface latent heat flux from ERA5 is higher than ~~in~~ the one from ERA-I ~~in only~~ only in a few areas, such as the central CONUS, eastern Australia, and eastern Europe. For the surface sensible heat flux, differences between ERA5 and ERA-I are clearly defined, with substantially higher values in the equatorial forests and lower values in (semi-)arid regions in the case of ERA5.

4 Conclusions

This study evaluated the surface energy partitioning over land in ECMWF's latest reanalysis ERA5 by assessing the quality of the surface latent heat flux, surface sensible heat flux, and Bowen ratio at different spatio-temporal scales and using different validation approaches. Results were also compared with the predecessor ERA-I for reference. Different in situ validation data sets – including eddy-covariance, river discharge, and balloon sounding data – were used to ~~directly~~ validate the reanalysis fields, and GLEAM and CLASS4GL were adopted as modelling tools to evaluate the surface energy partitioning in both reanalyses.

In a first experiment, the turbulent fluxes and the Bowen ratio from the reanalyses were directly compared against eddy-covariance measurements from the FLUXNET 2015 data set. The analysis revealed that ERA5 performed consistently better than ERA-I for all variables analysed, both at daily and sub-daily temporal resolutions, resulting in lower MAD and higher R against in situ data. The differences were most clear when anomaly time series were analysed, indicating that – although statistics also improved in case of the raw time series – ERA5 is substantially better capturing the response of surface energy partitioning to specific meteorological events. As one of the key changes in ERA5 is the use of the state-of-the-art H-TESSEL land-surface model and improvements in the land-surface data assimilation system, an important part of the improvements may be attributed to the improved land parameterisation. However, a validation of some key meteorological variables against

in situ measurements also showed better quality of these parameters from ERA5 than from ERA-I. These results were largely confirmed by an experiment where GLEAM was forced with meteorological fields retrieved from both reanalyses, showing a higher quality of the output based on ERA5 forcing data. Finally, although ERA5 did not seem to perform particularly better than ERA-I in specific climates, it was shown that GLEAM forced with ERA5 meteorology performed better than ERA5 in terms of estimating the surface latent heat flux in warm and dry regimes, indicating possible shortcomings in the land-surface scheme to capture the response of surface energy partitioning to heat and drought stress in ERA5.

In a second experiment, catchment-scale turbulent fluxes derived using discharge, precipitation, net radiation, and ground heat flux data were used to verify the bias in the annual turbulent fluxes from ERA-I and ERA5. Here, a substantial overestimation of the surface latent heat flux from ERA-I became evident. On the other hand, the surface sensible heat flux appeared generally underestimated. While the biases in ERA5 for the surface latent heat flux were found to be lower – a strong reduction was found along the east coast of the CONUS and in the south of Europe – a general tendency to overestimate the latent heat flux still remains in ERA5. In case of the surface sensible heat flux on the other hand, the sign of the bias reversed (i.e. in ERA5 the flux tends to be overestimated) and increased in absolute value.

A better quality of the surface energy partitioning in ERA5 was also confirmed by an experiment where CLASS4GL was forced with ~~surface fluxes~~ the evaporative fraction from ERA-I and ERA5~~, and outputs~~. Simulations of the diurnal evolution of the ABL were validated against a global archive of balloon soundings. CLASS4GL forced with ERA5 showed an overall better skill for simulating the diurnal boundary layer dynamics than ~~CLASS4GL~~ when forced with ERA-I. Especially in reproducing the tendencies of specific humidity, CLASS4GL seemed to strongly benefit from the seemingly better surface energy partitioning in ERA5, resulting in a substantially lower bias~~of the modelled variable~~. The latter could be attributed to the lower bias in the surface latent heat flux in ERA5 than in ERA-I. Since ~~ERA5 forced~~ ERA5-forced experiments better explained the global variability of the boundary layer dynamics, this experiment confirmed the overall better surface energy partitioning in ERA5 than in ERA-I, ~~which is~~ in line with the other independent experiments presented ~~in this paper~~here.

Finally, the global patterns of turbulent fluxes and Bowen ratio were analysed, and the globally-averaged magnitude of the fluxes was compared with values reported in literature. While the spatial patterns are realistic in both data sets, and align with the expectations from the major hydro-climatological regions, the substantial overestimation of the surface latent heat flux in both reanalyses emerged ~~once~~ again. However, the magnitude of the surface latent heat flux was found to be about 5% lower in ERA5 than in ERA-I, pointing towards the reduction of the bias, while the surface sensible heat flux only increased approximately 1%. The main reductions in the surface latent heat flux were found in regions that ~~have~~ had previously been highlighted in literature as hotspots of overestimation in ERA-I, such as the south of Europe, the Sahel, India, large parts of South America, and the east coast of the CONUS.

In ~~summary,~~ this paper, a variety of methods and data sets were used to evaluate the quality of the turbulent fluxes (and near-surface meteorology) from ERA5. As discussed throughout the manuscript, all techniques and reference data sets come with their own uncertainties and are derived based on different assumptions leading to potential flaws in the analyses presented in this paper. Eddy-covariance sites in the FLUXNET data set are not uniformly distributed across the globe, neither are the discharge measurements and balloon soundings used in this study. Therefore, conclusions should not be extrapolated to regions

that are under-represented in these data sets. In addition, the quality of each reference data set is affected by measurement errors and uncertainties introduced by assumptions made during the processing. Finally, both GLEAM and CLASS4GL are models and cannot be treated as ground truth as their estimates are impacted by uncertainties introduced by the model structure and parameterisation, as well as their inputs. Nevertheless, most analyses point into the direction of improvements from ERA-I to ERA5, irrespective of the validation technique or reference data set used, giving confidence to the conclusions draw in this study. In summary, it can be concluded that – based on the validation data and tools used in this study ~~,~~ – the quality of the turbulent fluxes (and near-surface meteorology) from ERA5 ~~shows a higher accuracy upon its predecessor~~has been improved. Although biases (especially in the surface latent heat flux) still prevail, changes in the IFS from ERA-I to ERA5, and improvements in the observational data sets that are assimilated into the models, have thus generally resulted in ~~enhancements in the right direction~~a higher-quality surface energy partitioning in the reanalysis.

Code and data availability. All data sets used in this study can be freely accessed from their respective repositories after registration. ERA-I data were downloaded from the ECMWF web page (https://apps.ecmwf.int/datasets/data/), ERA5 data were retrieved from the Copernicus Climate Data Store (https://cds.climate.copernicus.eu/), GLEAM data were accessed from https://www.gleam.eu/, GRDC discharge data can be downloaded from https://www.bafg.de/GRDC/EN/02_srvcs/21_tmsrs/riverdischarge_node.html, the FLUXNET2015 Tier2 data set can be accessed from the FLUXNET data portal at https://fluxnet.fluxdata.org/data/fluxnet2015-dataset/, input data for CLASS4GL is available at https://www.CLASS4GL.eu/, and the output of CLASS4GL is available upon request. The source code of CLASS4GL can be accessed at https://www.CLASS4GL.eu/.

Table 1. Averaged metrics and their confidence interval of surface energy partitioning from ERA5, ERA-I, GLEAM+ERA5, and GLEAM+ERA-I across the FLUXNET 2015 data set.

		$\lambda\rho E$ (3h) W m^{-2}	$\lambda\rho E$ (24h) W m^{-2}	H (3h) W m^{-2}	H (24h) W m^{-2}	β (24h) –
MD	ERA5	9.27 (±0.080)	8.49 (±0.178)	-2.60 (±0.010)	-2.99 (±0.140)	-0.56 (±0.013)
	ERA-I	11.12 (±0.079)	10.29 (±0.180)	-3.38 (±0.099)	-3.66 (±0.147)	-0.69 (±0.012)
	GLEAM+ERA5	n.a.	-3.27 (±0.176)	n.a.	-5.83 (±0.153)	-0.25 (±0.014)
	GLEAM+ERA-I	n.a.	-3.76 (±0.179)	n.a.	-10.14 (±0.158)	-0.39 (±0.014)
R	ERA5	0.34 (±0.002)	0.41 (±0.005)	0.46 (±0.002)	0.50 (±0.004)	0.39 (±0.006)
	ERA-I	0.31 (±0.002)	0.39 (±0.005)	0.42 (±0.002)	0.45 (±0.004)	0.36 (±0.006)
	GLEAM+ERA5	n.a.	0.35 (±0.005)	n.a.	0.45 (±0.005)	0.39 (±0.006)
	GLEAM+ERA-I	n.a.	0.32 (±0.005)	n.a.	0.46 (±0.005)	0.40 (±0.007)

Appendix A: ~~Supplementary figures~~

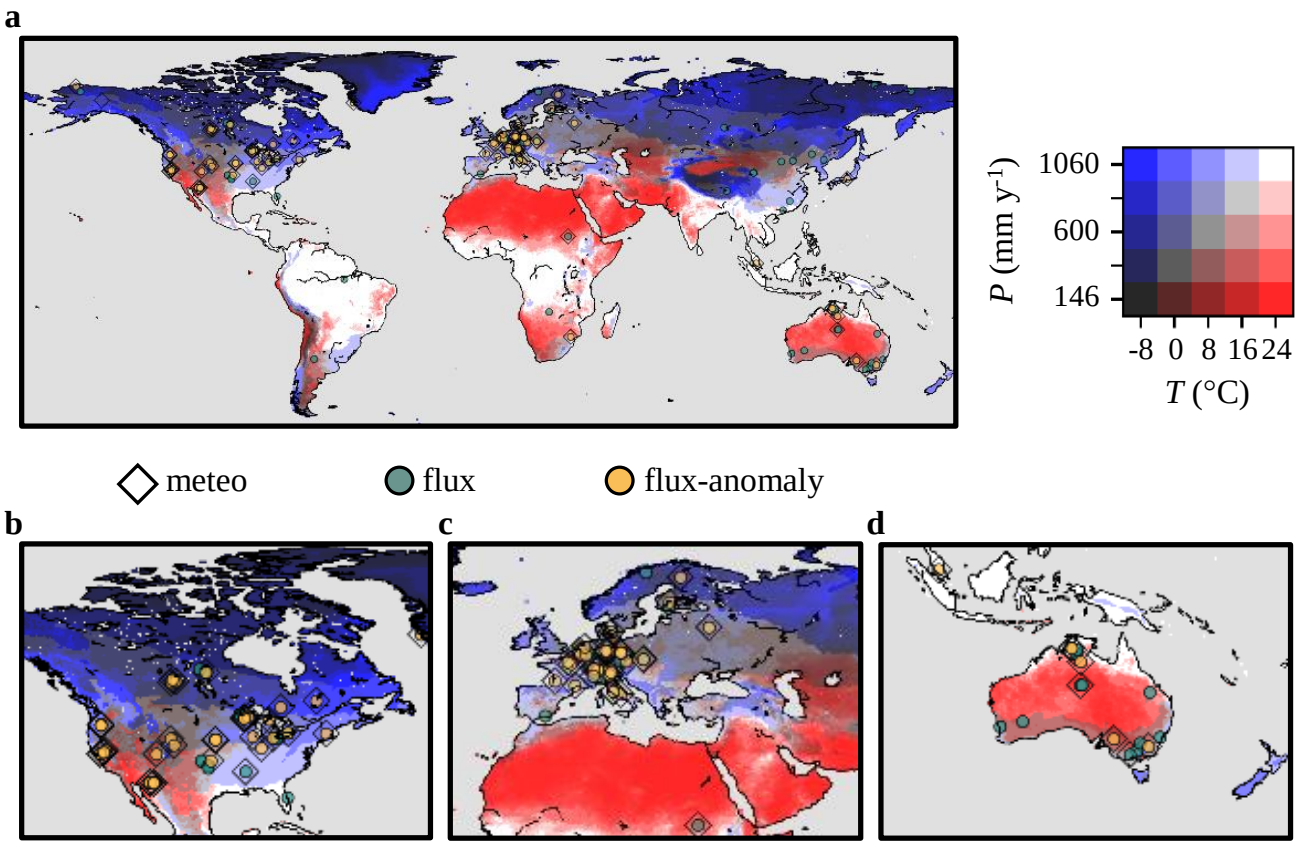

Figure 1. **(a)** Location of the selected eddy-covariance sites. **(b)–(c)** show a detailed view of the sites across the CONUS **(b)**, Europe **(c)**, and Autralia **(d)**. Sites with a record length of less than 5 years (i.e. where no anomalies are calculated) are plotted in green and sites with a record length of more than 5 years (i.e. where anomalies are calculated) are plotted in yellow. Sites where measurements of meteorological data are also available are indicated with a diamond. The background provides information on the climatological mean temperature and precipitation derived from ERA5 (1983–2018).

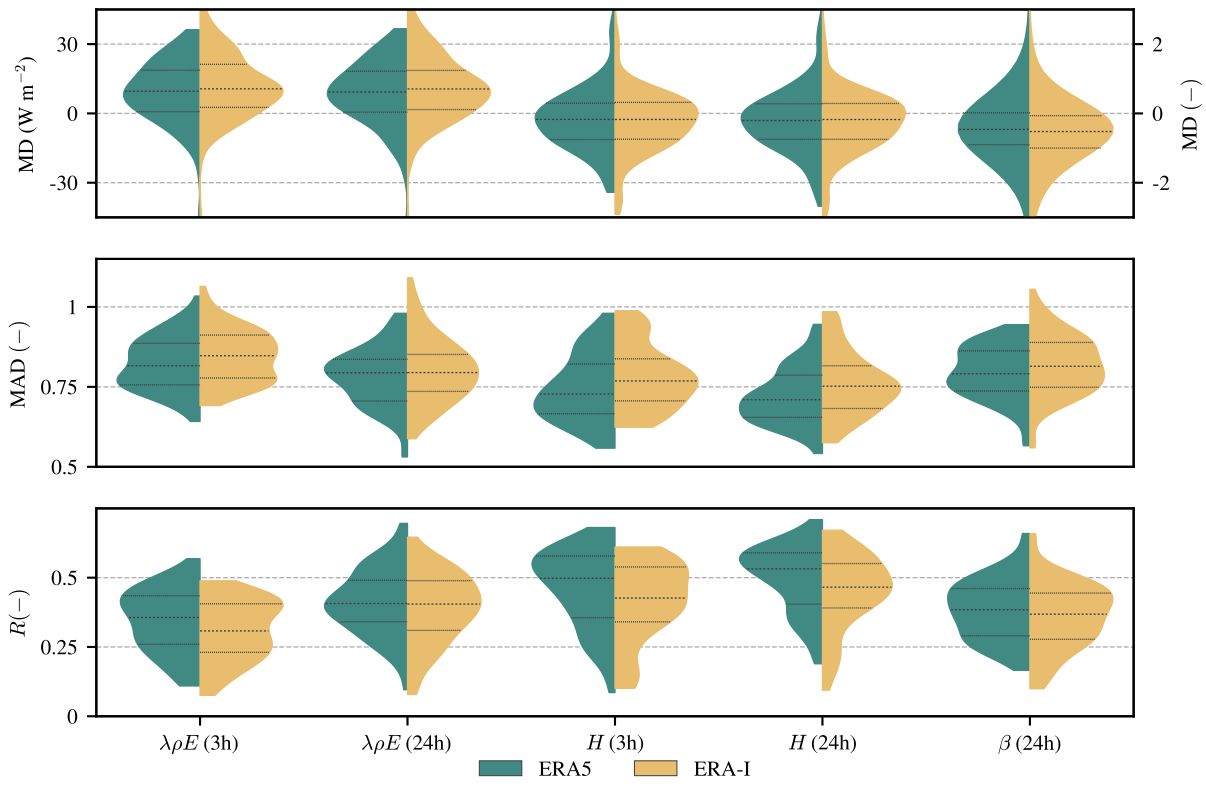

Figure 2. Violin plots of temporal validation statistics of the surface latent heat flux ($\lambda\rho E$), surface sensible heat flux (H), and Bowen ratio (β) from ERA5 (green) and ERA-I (yellow). Statistics are calculated against in situ eddy-covariance measurements at both 3-hourly and daily temporal resolutions. Violin plots represent the distribution of the individual validation statistics with indication of the median and inter-quartile range, and are calculated using a kernel density estimation approach. Statistics include the Mean Difference (MD, raw in situ time series from 143 sites as reference), Mean Absolute Difference (MAD, anomaly in situ time series from 77 sites as reference), and the Pearson correlation coefficient (R, anomaly in situ time series from 77 sites as reference). The distribution of the MD of β is plotted on the right y-axis.

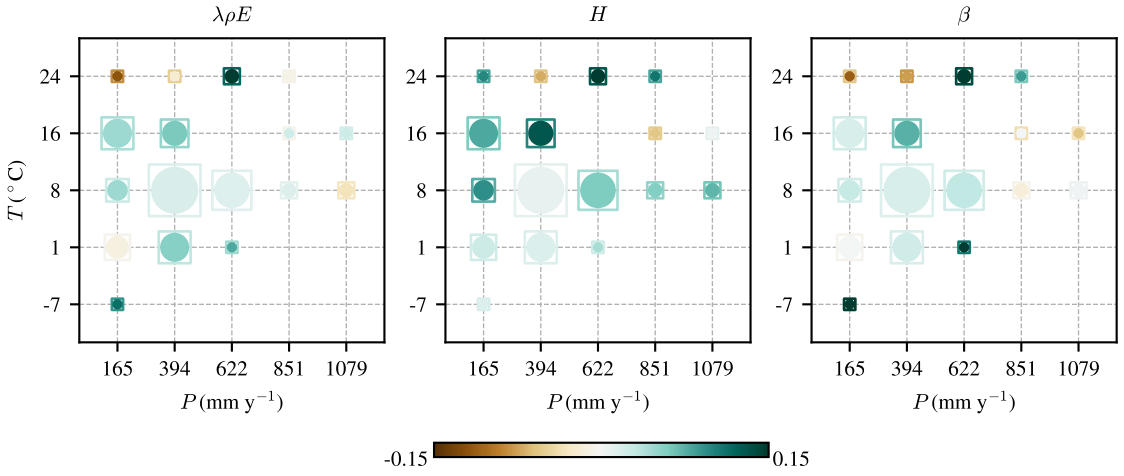

Figure 3. Difference between temporal validation statistics of the surface latent heat flux ($\lambda\rho E$), surface sensible heat flux (H), and Bowen ratio (β) from ERA5 and ERA-I grouped as a function of precipitation rate (P) and near-surface air temperature (T) calculated at the in situ site. Statistics are calculated against in situ eddy-covariance measurements at daily resolution and then averaged across the sites within each group. Statistics include the Mean Absolute Difference (MAD, anomaly in situ time series from 77 sites as reference) and the Pearson correlation coefficient (R, anomaly in situ time series from 77 sites as reference). Circles show the R from ERA5 minus the one from ERA-I, while squares show the MAD from ERA-I minus the one from ERA5; hence, green colors represent better statistics for ERA5 compared to ERA-I. The size of the symbols relates to the number of in situ sites per group.

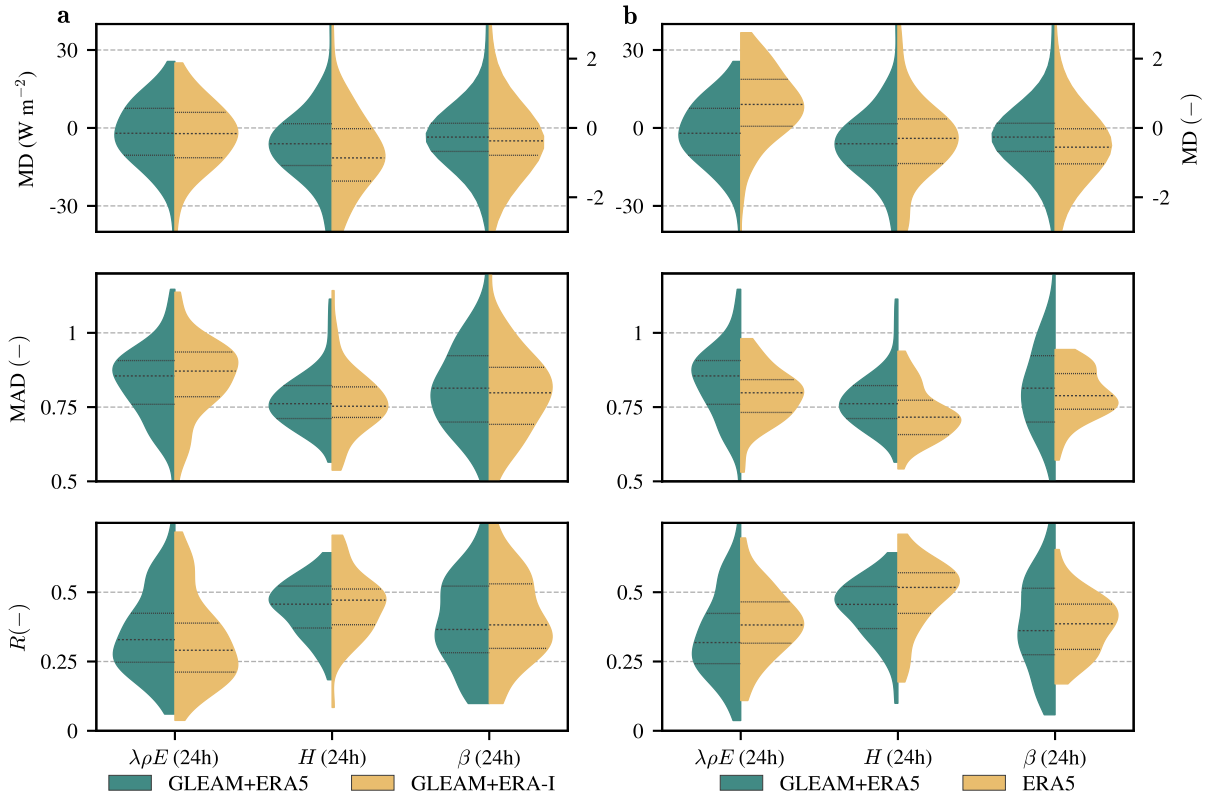

Figure 4. Violin plots of temporal validation statistics of the surface latent heat flux ($\lambda\rho E$), surface sensible heat flux (H), and Bowen ratio (β) from GLEAM+ERA5, GLEAM+ERA-I, and ERA5. **(a)** Compares the violin plots from GLEAM+ERA5 and GLEAM+ERA-I and **(b)** compares the violin plots from GLEAM+ERA5 and ERA5. Statistics are calculated against in situ eddy-covariance measurements at daily temporal resolution. Violin plots represent the distribution of the individual validation statistics with indication of the median and inter-quartile range, and are calculated using a kernel density estimation approach. Statistics include the Mean Difference (MD, raw in situ time series from 143 sites as reference), Mean Absolute Difference (MAD, anomaly in situ time series from 77 sites as reference), and the Pearson correlation coefficient (R, anomaly in situ time series from 77 sites as reference). The distribution of the MD of β is plotted on the right y-axis.

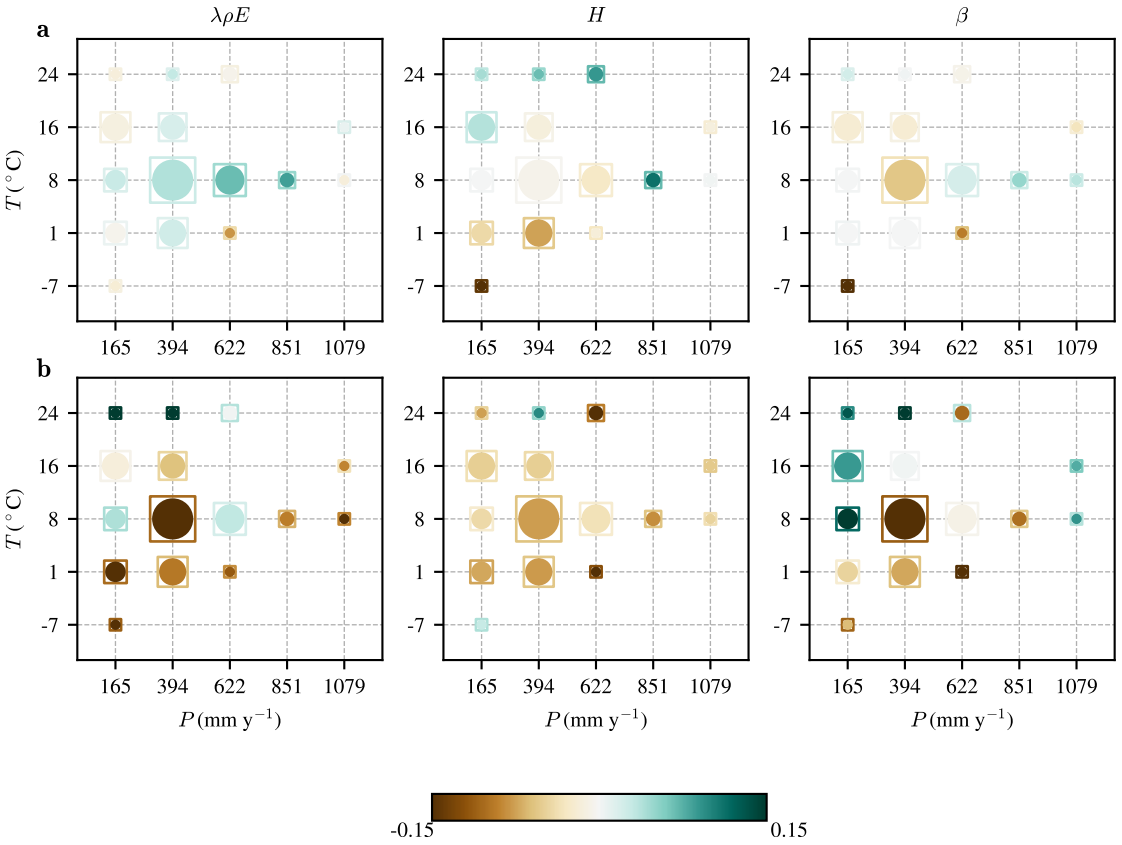

Figure 5. Difference between temporal validation statistics of the surface latent heat flux ($\lambda\rho E$), surface sensible heat flux (H), and Bowen ratio (β) from GLEAM+ERA5, GLEAM+ERA-I, and ERA5 grouped as a function of precipitation rate (P) and near-surface air temperature (T) calculated at the in situ site. **(a)** Compares the statistics from GLEAM+ERA5 and GLEAM+ERA-I and **(b)** compares the statistics from GLEAM+ERA5 and ERA5. Statistics are calculated against in situ eddy-covariance measurements at daily resolution and then averaged across the sites within each group. Statistics include the Mean Absolute Difference (MAD, anomaly in situ time series from 77 sites as reference) and the Pearson correlation coefficient (R, anomaly in situ time series from 77 sites as reference). In **(a)** circles show the R from GLEAM+ERA5 minus the one from GLEAM+ERA-I, while squares show the MAD from GLEAM+ERA-I minus the one from GLEAM+ERA5; hence, green colors represent better statistics for GLEAM+ERA5 compared to GLEAM+ERA-I. In **(b)**, statistics from GLEAM+ERA-I are replaced by ERA5; hence, green colors represent better statistics for GLEAM+ERA5 compared to ERA5. The size of the symbols relates to the number of in situ sites per group.

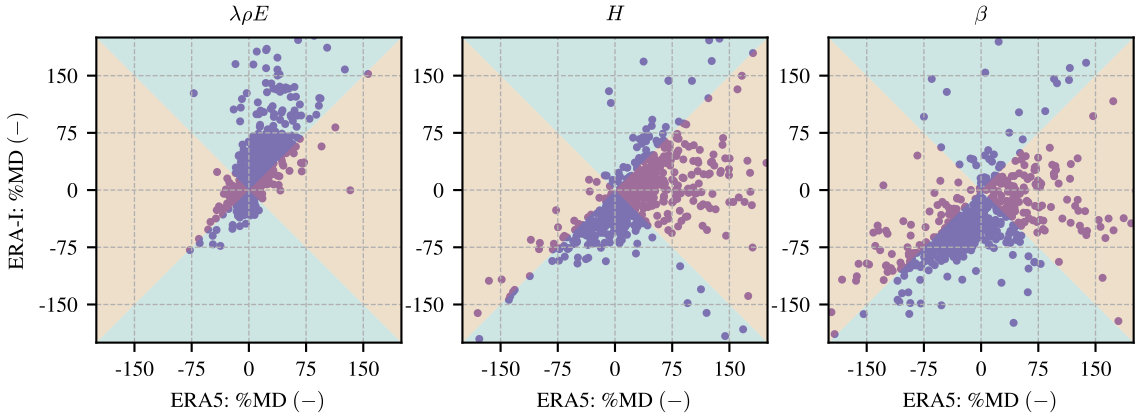

Figure 6. Scatter plot of the bias of the surface latent heat flux ($\lambda\rho E$), surface sensible heat flux (H), and Bowen ratio (β) from ERA-I versus ERA5. The bias is calculated against catchment-scale estimates of the fluxes derived using discharge data (Eqs. 1–3) and is assessed by the percentage Mean Difference (%MD, raw time series from 707 catchments as reference). The green area indicates points where the bias in ERA5 is better than in ERA-I, and vice versa for the brown area.

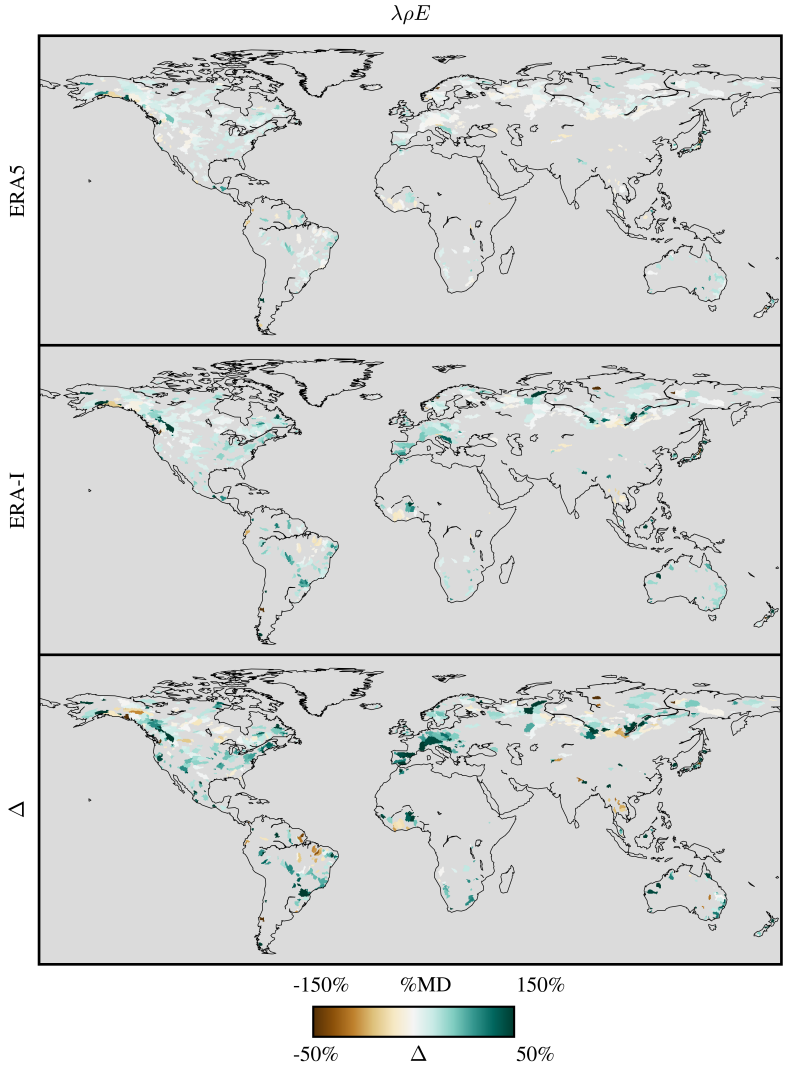

Figure 7. Maps of the bias of the surface latent heat flux ($\lambda\rho E$) from ERA5 and ERA-I. The bias is calculated against catchment-scale estimates of the fluxes derived using discharge data (Eqs. 1–3) and is assessed by the percentage Mean Difference (%MD, raw time series from 707 catchments as reference). The bottom map represents the difference (Δ) between the absolute bias in ERA-I and ERA5; hence, green colors represent lower bias in ERA5 than in ERA-I.

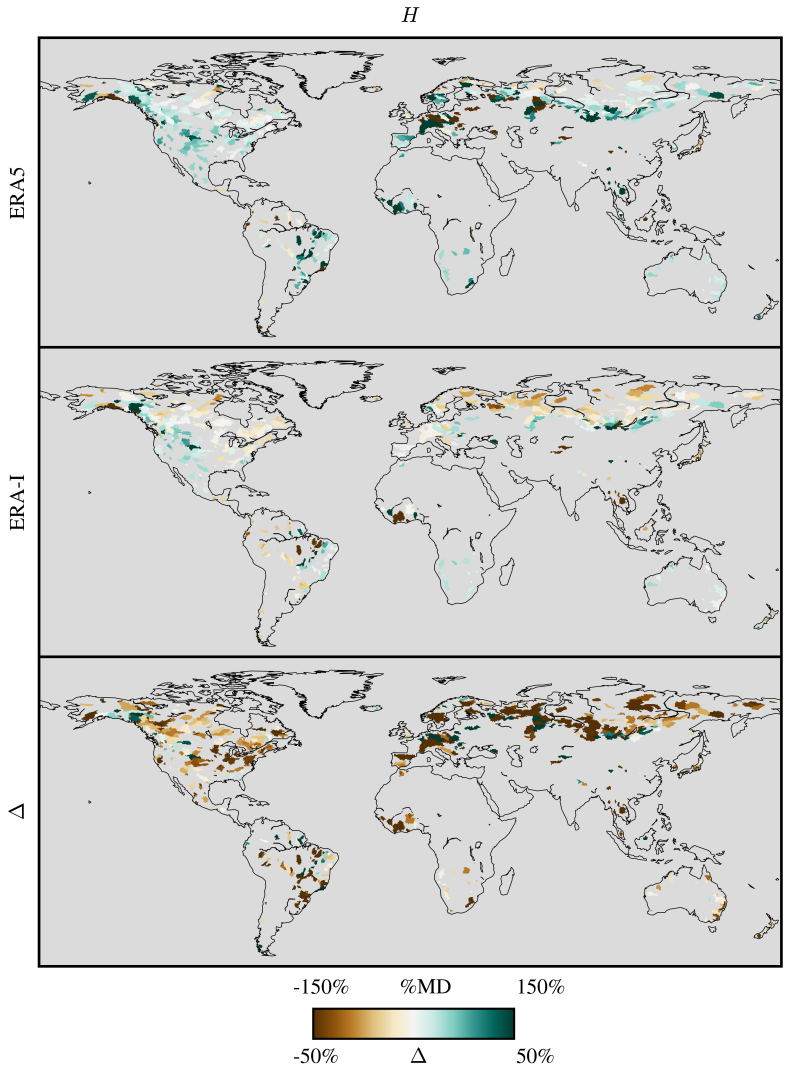

Figure 8. Like Fig. 7, but for the surface sensible heat flux (*H*).

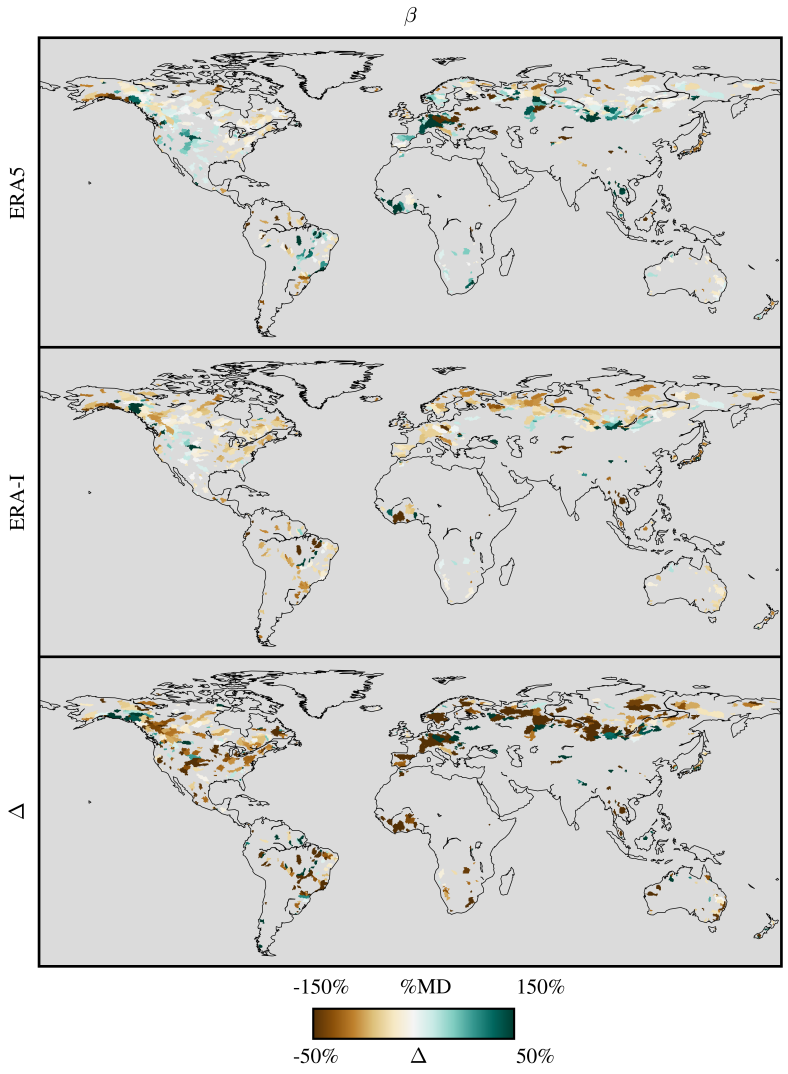

Figure 9. Like Fig. 7, but for the Bowen ratio (β).

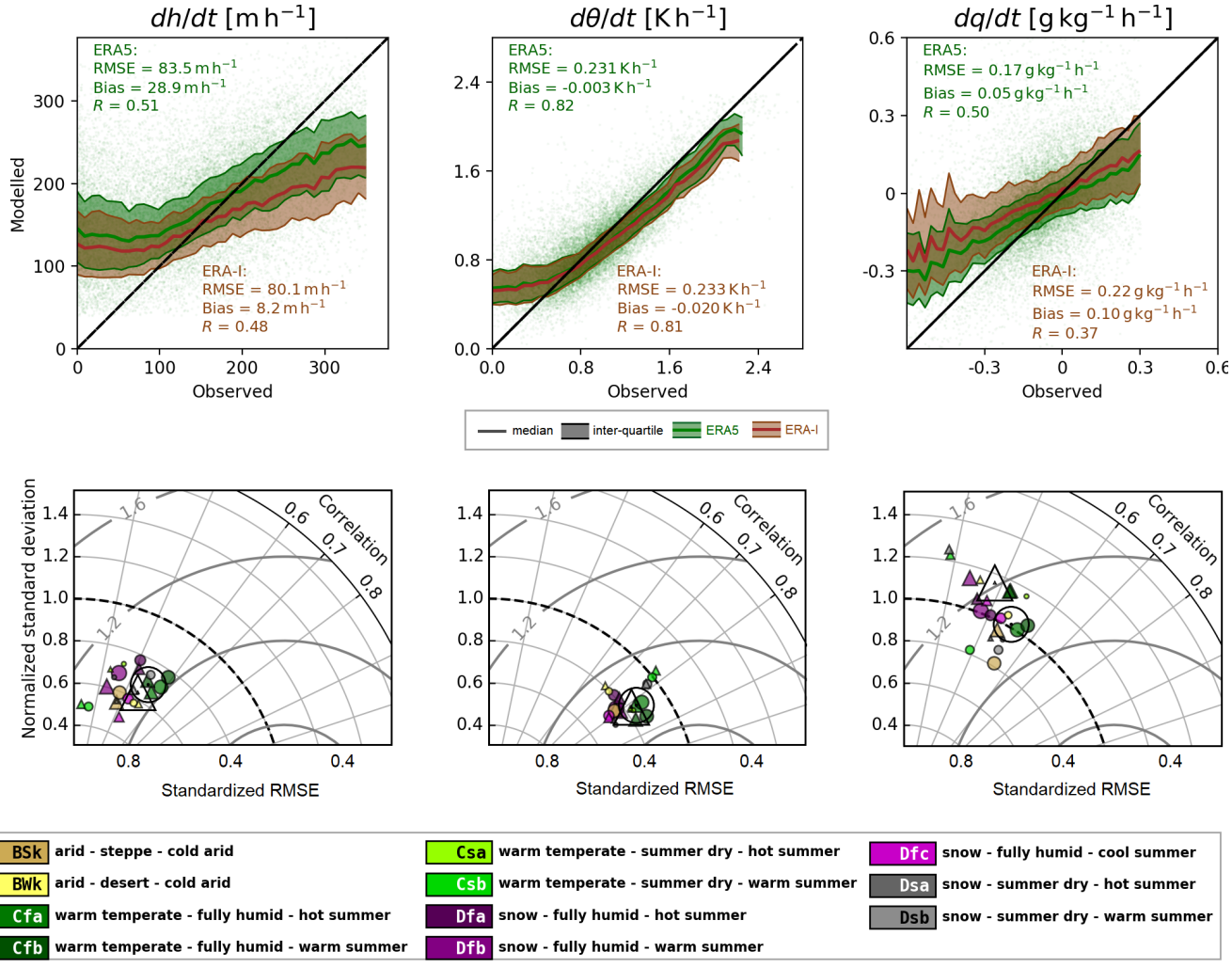

Figure 10. Skill of CLASS4GL at reproducing diurnal changes in ABL properties when forced with surface evaporative fractions from ERA5 versus ERA-I. Shown are the tendencies of the mixed-layer height (dh/dt), potential temperature ($d\theta/dt$), and specific humidity (dq/dt), which are assessed by comparison of model simulations against the IGRA sounding data between 1981 and 2015. The first row shows modeled versus observed data points, and the corresponding median and inter-quartile range of the simulations in solid lines, where green represents ERA5 and brown ERA-I. The 1–1 line is shown as a black line for reference. The bottom row illustrates the skill of the ABL simulations when forced with ERA5 (circles) versus ERA-I (triangles) in the form of Taylor plots. The transparent symbols show the overall performance of 18000 sounding pairs from 121 stations, whereas the colored symbols indicate the performance per Köppen-Geiger climate class and for which the size is proportional to the number of sounding pairs.

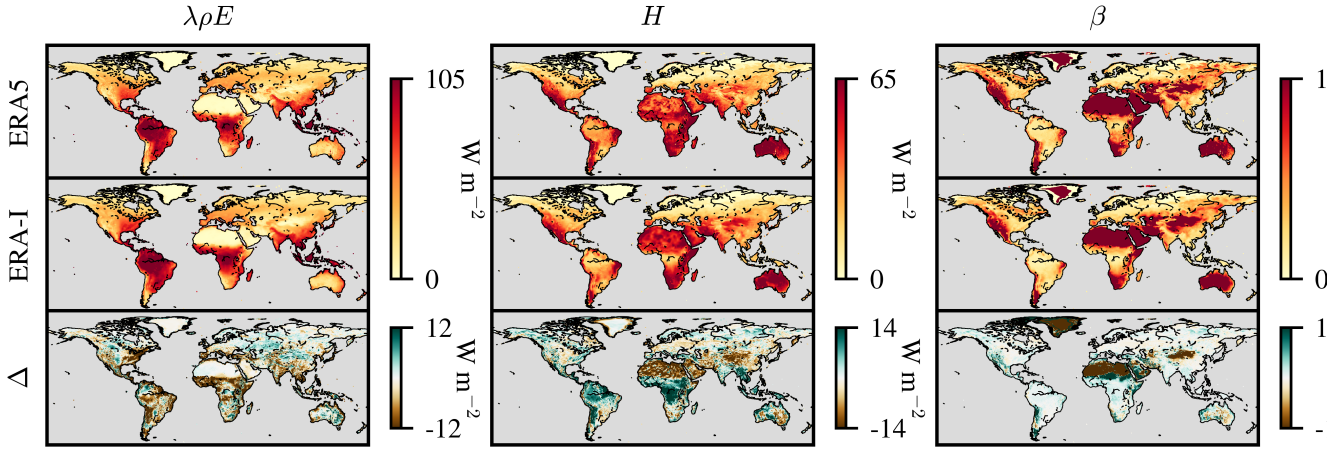

Figure 11. Maps of the multi-annual average of surface latent heat flux ($\lambda\rho E$, W m^{-2}), surface sensible heat flux (H, W m^{-2}), and Bowen ratio (β) from ERA5 and ERA-I. In the last row, Δ presents the difference between ERA5 and ERA-I; hence, green colors represent higher values in ERA5 compared to ERA-I.

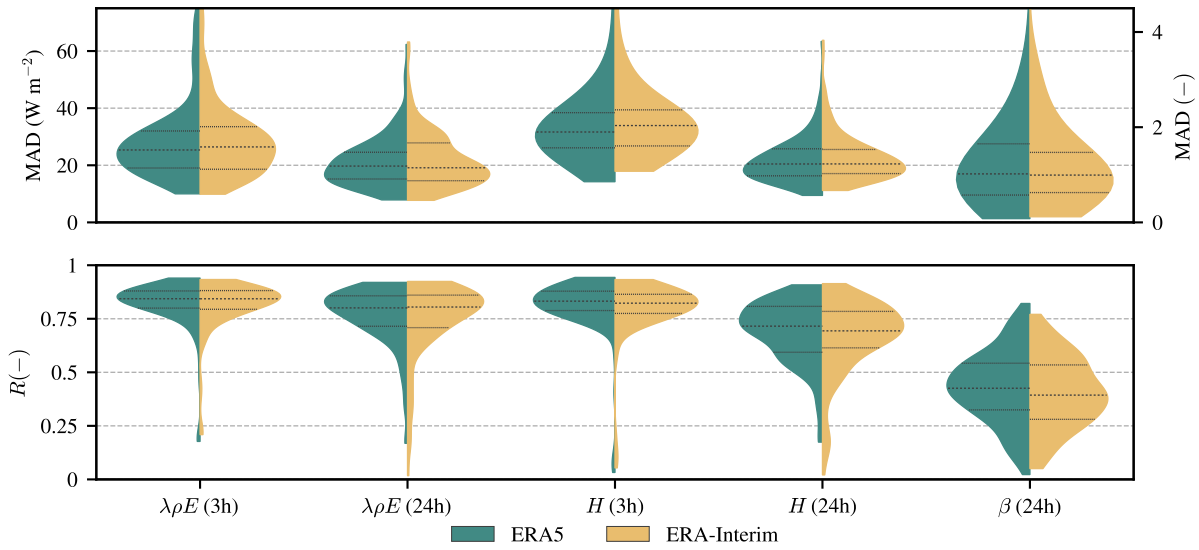

Figure A1. Violin plots of temporal validation statistics of the surface latent heat flux ($\lambda\rho E$), surface sensible heat flux (H), and Bowen ratio (β) from ERA5 (green) and ERA-I (~~orange~~yellow). Statistics are calculated against in situ eddy-covariance measurements at both 3-hourly and daily temporal resolutions. ~~Violins~~ Violin plots represent the distribution of the individual validation statistics with indication of the median and inter-quartile range, and are calculated using a kernel density estimation approach. Statistics include the Mean Absolute Difference (MAD, raw in situ time series from 143 sites as reference) and the Pearson correlation coefficient (R, raw in situ time series from 143 sites as reference). ~~For MAD, the~~ The distribution of the MAD of β is plotted on the right y-axis.

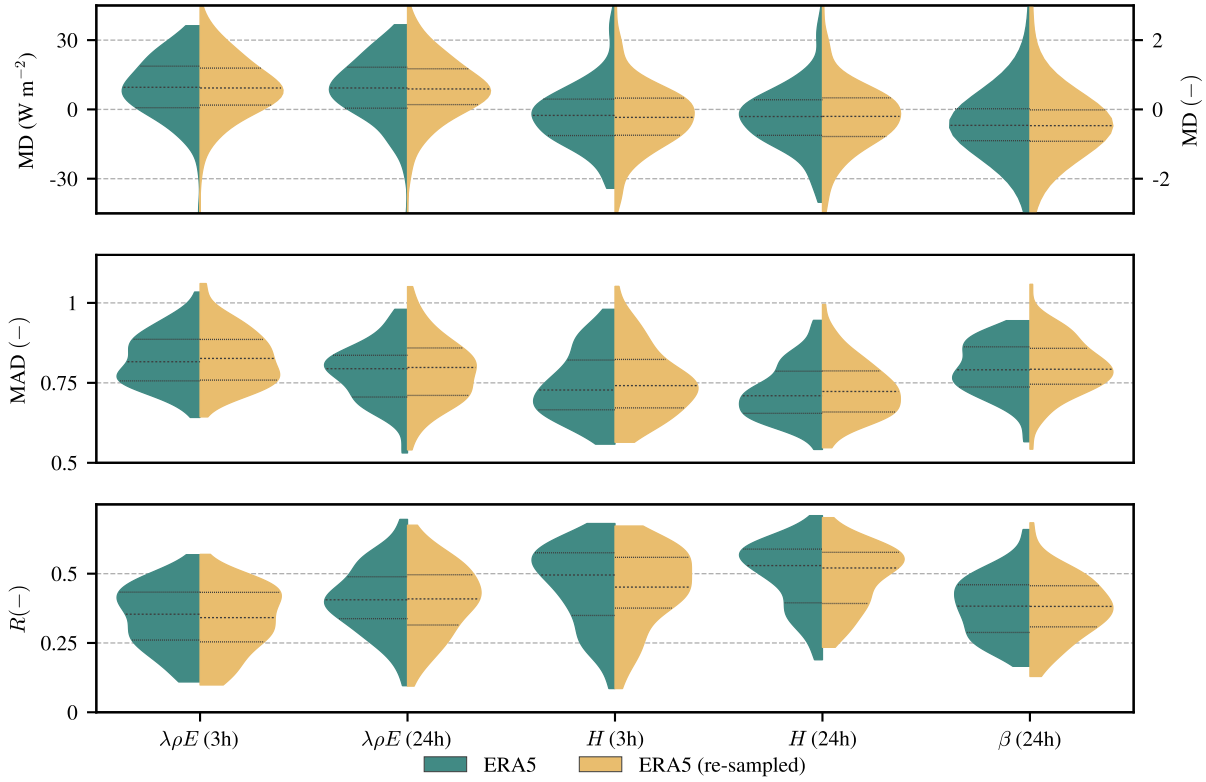

Figure A2. Violin plots of temporal validation statistics of the surface latent heat flux ($\lambda\rho E$), surface sensible heat flux (H), and Bowen ratio (β) from ERA5 (green) and ERA5 linearly re-sampled to the spatial grid of ERA-I (~~orange~~yellow). Statistics are calculated against in situ eddy-covariance measurements at both 3-hourly and daily temporal resolutions. ~~Violins~~ Violin plots represent the distribution of the individual validation statistics with indication of the median and inter-quartile range and are calculated using a kernel density estimation approach. Statistics include the Mean Difference (MD, raw in situ time series from 143 sites as reference), Mean Absolute Difference (MAD, anomaly in situ time series from 77 sites as reference), and the Pearson correlation coefficient (R, anomaly in situ time series from 77 sites as reference). ~~For MD, the~~ The distribution of the MD of β is plotted on the right y-axis.

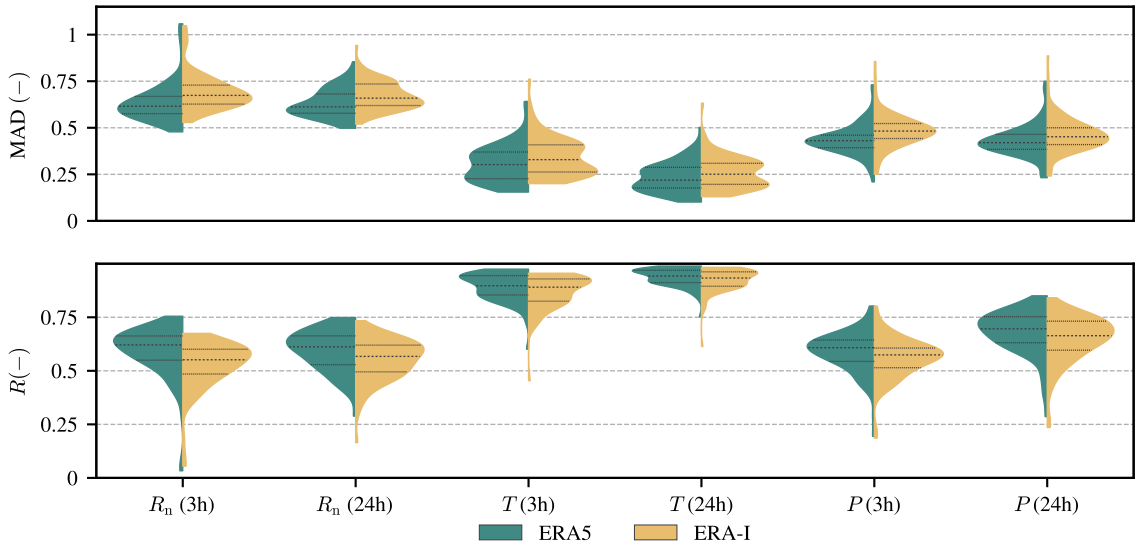

Figure A3. Violin plots of temporal validation statistics of the surface net radiation (R_n), 2-meter air temperature (T), and precipitation rate (P) from ERA5 (green) and ERA-I (~~orange~~yellow). Statistics are calculated against in situ eddy-covariance measurements at both 3-hourly and daily temporal resolutions. ~~Violins~~ Violin plots represent the distribution of the individual validation statistics with indication of the median and inter-quartile range, and are calculated using a kernel density estimation approach. Statistics include the Mean Absolute Difference (MAD, anomaly in situ time series from 83 sites as reference) and the Pearson correlation coefficient (R, anomaly in situ time series from 83 sites as reference).

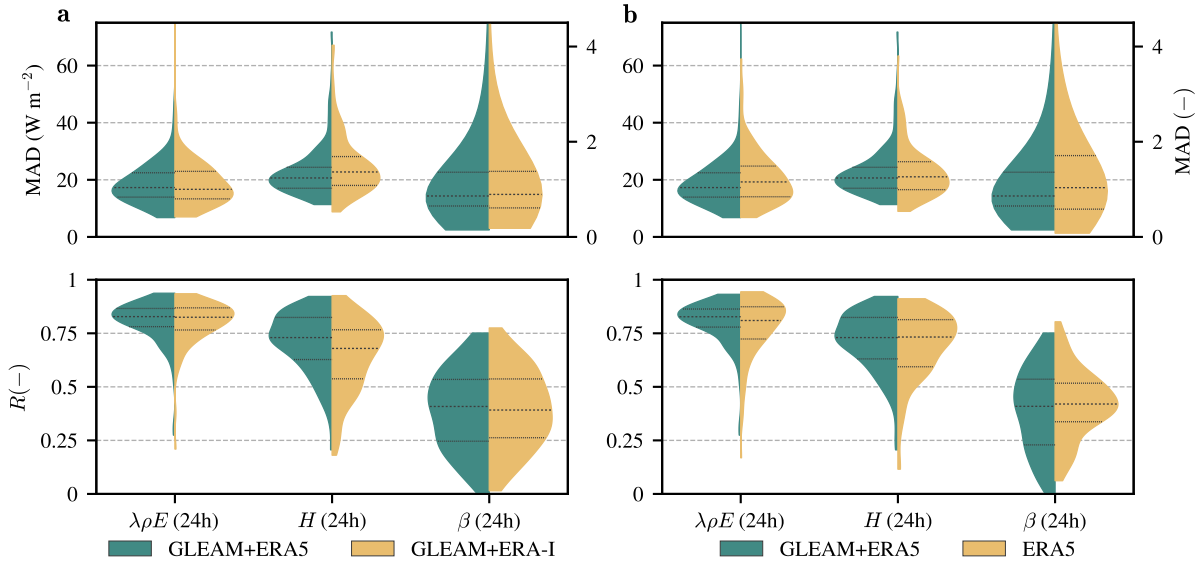

Figure A4. Violin plots of temporal validation statistics of the surface latent heat flux ($\lambda\rho E$), surface sensible heat flux (H), and Bowen ratio (β) from GLEAM+ERA5, GLEAM+ERA-I, and ERA5. **(a)** Compares the ~~violins~~ violin plots from GLEAM+ERA5 and GLEAM+ERA-I and **(b)** directly compares the ~~violins~~ violin plots from GLEAM+ERA5 and ERA5. Statistics are calculated against in situ eddy-covariance measurements at daily temporal resolution. ~~Violins~~ Violin plots represent the distribution of the individual validation statistics with indication of the median and inter-quartile range and are calculated using a kernel density estimation approach. Statistics include the Mean Absolute Difference (MAD, raw in situ time series from 143 sites as reference) and the Pearson correlation coefficient (R, raw in situ time series from 143 sites as reference). ~~For MAD, the~~ The distribution of the MAD of β is plotted on the right y-axis.

Author contributions. B.M. and D.G.M. conceived the study and designed the lay-out. B.M. processed the reanalyses data, eddy-covariance data, and performed the simulations with GLEAM. D.S. processed the GRDC discharge data. H.W. processed the balloon sounding data and performed the CLASS4GL simulations. B.M. and D.G.M. designed the lay-out of the manuscript and B.M. led the writing. All authors have been involved in interpreting the results, discussing the findings, and editing the manuscript.

Competing interests. The authors declare no competing interests.

Acknowledgements. This work was partly funded by the Belgian Science Policy Office through the ET-Sense project (SR/02/377). D.G.M., H.W., and D.S. acknowledge support from the European Research Council (ERC) under grant agreement n° 715254 (DRY-2-DRY). This work used eddy-covariance data acquired and shared by the FLUXNET community, including these networks: AmeriFlux, AfriFlux, AsiaFlux, CarboAfrica, CarboEuropeIP, CarboItaly, CarboMont, ChinaFlux, Fluxnet-Canada, GreenGrass, ICOS, KoFlux, LBA, NECC, OzFlux-TERN, TCOS-Siberia, and USCCC. The FLUXNET eddy covariance data processing and harmonization was carried out by the ICOS Ecosystem Thematic Center, AmeriFlux Management Project and Fluxdata project of FLUXNET, with the support of CDIAC, and the OzFlux, ChinaFlux and AsiaFlux offices. The authors also acknowledge Instituto Nacional Technologica Agropecuaria for making the eddy-covariance data of AR-Vir publically available.

References

Albergel, C., Balsamo, G., De Rosnay, P., Muñoz-Sabater, J., and Boussetta, S.: A bare ground evaporation revision in the ECMWF land-surface scheme: evaluation of its impact using ground soil moisture and satellite microwave data, Hydrology and Earth System Sciences, 16, 3607–3620, https://doi.org/10.5194/hess-16-3607-2012, 2012.

Albergel, C., Dutra, E., Munier, S., Calvet, J. C., Muñoz Sabater, J., De Rosnay, P., and Balsamo, G.: ERA-5 and ERA-Interim driven ISBA land surface model simulations: which one performs better?, Hydrology and Earth System Sciences, 22, 3515–3532, https://doi.org/10.5194/hess-22-3515-2018, 2018.

Baldocchi, D., Falge, E., Gu, L. H., Olson, R., Hollinger, D., Running, S., Anthoni, P., Bernhofer, C., Davis, K., Evans, R., Fuentes, J., Goldstein, A., Katul, G., Law, B., Lee, X. H., Malhi, Y., Meyers, T., Munger, W., Oechel, W., 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, https://doi.org/10.1175/1520-0477(2001)082<2415:FANTTS>2.3.CO;2, 2001.

Balsamo, G., Beljaars, A., Scipal, K., Viterbo, P., van den Hurk, B., Hirschi, M., and Betts, A. K.: A revised hydrology for the ECMWF model: verification from field site to terrestrial water storage and impact in the integrated forecast system, Journal of Hydrometeorology, 10, 623–643, https://doi.org/10.1175/2008jhm1068.1, 2008.

Balsamo, G., Albergel, C., Beljaars, A., Boussetta, S., Brun, E., Cloke, H., Dee, D., Dutra, E., Munõz-Sabater, J., Pappenberger, F., De Rosnay, P., Stockdale, T., and Vitart, F.: ERA-Interim/Land: a global land surface reanalysis data set, Hydrology and Earth System Sciences, 19, 389–407, https://doi.org/10.5194/hess-19-389-2015, 2015.

Beck, H. E., Wood, E. F., Pan, M., Fisher, C. K., Miralles, D. G., van Dijk, A. I. J. M., McVicar, T. R., and Adler, R. F.: MSWEP V2 Global 3-Hourly 0.1° Precipitation: Methodology and Quantitative Assessment, Bulletin of the American Meteorological Society, 100, 473–500, https://doi.org/10.1175/BAMS-D-17-0138.1, 2019.

Beer, C., Reichstein, M., Tomelleri, E., Ciais, P., Jung, M., Carvalhais, N., Rödenbeck, C., Arain, M. A., Baldocchi, D., Bonan, G. B., Bondeau, A., Cescatti, A., Lasslop, G., Lindroth, A., Lomas, M., Luyssaert, S., Margolis, H., Oleson, K. W., Roupsard, O., Veenendaal, E., Viovy, N., Williams, C., Woodward, F. I., and Papale, D.: Terrestrial gross carbon dioxide uptake: global distribution and covariation with climate, Science, 329, 834–838, https://doi.org/10.1126/science.1184984, 2010.

Berg, A. and Sheffield, J.: Soil moisture–evapotranspiration coupling in CMIP5 models: relationship with simulated climate and projections, Journal of Climate, 31, 4865–4878, https://doi.org/10.1175/JCLI-D-17-0757.1, 2018.

Brunamonti, S., Füzér, L., Jorge, T., Poltera, Y., Oelsner, P., Meier, S., Dirksen, R., Naja, M., Fadnavis, S., Karmacharya, J., Wienhold, F., Luo, B., Wernli, H., and Peter, T.: Water vapor in the Asian summer monsoon anticyclone: comparison of balloon-borne measurements and ECMWF data, Journal of Geophysical Research: Atmospheres, 124, JD030 000, https://doi.org/10.1029/2018jd030000, 2019.

De Lannoy, G. J. M. 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 estimation, Journal of Hydrometeorology, 17, 669–691, https://doi.org/10.1175/JHM-D-15-0037.1, 2016.

Decker, M., Brunke, M. A., Wang, Z., Sakaguchi, K., Zeng, X., and Bosilovich, M. G.: Evaluation of the reanalysis products from GSFC, NCEP, and ECMWF using flux tower observations, Journal of Climate, 25, 1916–1944, https://doi.org/10.1175/JCLI-D-11-00004.1, 2012.

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M. A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A. C. M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A. J., Haim-

berger, L., Healy, S. B., Hersbach, H., Hólm, E. V., Isaksen, L., Kållberg, P., Köhler, M., Matricardi, M., Mcnally, A. P., Monge-Sanz, B. M., Morcrette, J. J., Park, B. K., Peubey, C., de Rosnay, P., Tavolato, C., Thipaut, J. N., and Vitart, F.: The ERA-Interim reanalysis: configuration and performance of the data assimilation system, Quarterly Journal of the Royal Meteorological Society, 137, 553–597, https://doi.org/10.1002/qj.828, 2011.

Dehghani, A., Sarbishei, O., Glatard, T., and Shihab, E.: A quantitative comparison of overlapping and non-overlapping sliding windows for human activity recognition using inertial sensors, Sensors, 19, 5026, https://doi.org/10.3390/s19225026, 2019.

Dimiceli, C., Carroll, M., Sohlberg, R., Kim, D. H., Kelly, M., and Townshend, J. R. G.: MOD44B MODIS/Terra vegetation continuous fields yearly L3 global 250m SIN grid V006. NASA EOSDIS Land Processes DAAC [data set], doi: 10.5067/MODIS/MOD44B.006, 2015.

Dirmeyer, P. A., Chen, L., Wu, J., Shin, C.-S., Huang, B., Cash, B. A., Bosilovich, M. G., Mahanama, S., Koster, R. D., Santanello, J. A., Ek, M. B., Balsamo, G., Dutra, E., and Lawrence, D. M.: Verification of land–atmosphere coupling in forecast models, reanalyses, and land surface models using flux site observations, Journal of Hydrometeorology, 19, 375–392, https://doi.org/10.1175/jhm-d-17-0152.1, 2017.

Draper, C. S., Reichle, R. H., and Koster, R. D.: Assessment of MERRA-2 land surface energy flux estimates, Journal of Climate, 31, 671–691, https://doi.org/10.1175/JCLI-D-17-0121.1, 2018.

Durre, I., Vose, R. S., and Wuertz, D. B.: Overview of the Integrated Global Radiosonde Archive, Journal of Climate, 19, 53–68, 2006.

Gash, J. H. C.: An analytical model of rainfall interception by forests, Quarterly Journal of the Royal Meteorological Society, 105, 43–55, https://doi.org/10.1002/qj.49710544304, 1979.

Gelaro, R., McCarty, W., Suárez, M. J., Todling, R., Molod, A., Takacs, L., Randles, C. A., Darmenov, A., Bosilovich, M. G., Reichle, R., Wargan, K., Coy, L., Cullather, R., Draper, C., Akella, S., Buchard, V., Conaty, A., da Silva, A. M., Gu, W., Kim, G. K., Koster, R., Lucchesi, R., Merkova, D., Nielsen, J. E., Partyka, G., Pawson, S., Putman, W., Rienecker, M., Schubert, S. D., Sienkiewicz, M., and Zhao, B.: The Modern-Era Retrospective Analysis for research and applications, version 2 (MERRA-2), Journal of Climate, 30, 5419–5454, https://doi.org/10.1175/JCLI-D-16-0758.1, 2017.

Graham, R. M., Hudson, S. R., and Maturilli, M.: Improved performance of ERA5 in Arctic Gateway relative to four global atmospheric reanalyses, Geophysical Research Letters, 46, 6138–6147, https://doi.org/10.1029/2019GL082781, 2019.

Gruber, A., De Lannoy, G., Albergel, C., Al-Yaari, A., Brocca, L., Calvet, J.-C., Colliander, A., Cosh, M., Crow, W., Dorigo, W., Draper, C., Hirschi, M., Kerr, Y., Konings, A., Lahoz, W., McColl, K., Montzka, C., Muñoz-Sabater, J., Peng, J., Reichle, R., Richaume, P., Rüdiger, C., Scanlon, T., van der Schalie, R., Wigneron, J.-P., and Wagner, W.: Validation practices for satellite soil moisture retrievals: What are (the) errors?, Remote Sensing of Environment, 244, 111 806, https://doi.org/10.1016/j.rse.2020.111806, 2020.

Guillod, B. P., Orlowsky, B., Miralles, D. G., Teuling, A. J., and Seneviratne, S. I.: Reconciling spatial and temporal soil moisture effects on afternoon rainfall, Nature Communications, 6, 6443, https://doi.org/10.1038/ncomms7443, 2015.

Hersbach, H., de Rosnay, P., Bell, B., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Alonso-Balmaseda, M., Balsamo, G., Bechtold, P., Berrisford, P., Bidlot, J.-R., de Boisséson, E., Bonavita, M., Browne, P., Buizza, R., Dahlgren, P., Dee, D., Dragani, R., Diamantakis, M., Flemming, J., Forbes, R., Geer, A. J., Haiden, T., Hólm, E., Haimberger, L., Hogan, R., Horányi, A., Janiskova, M., Laloyaux, P., Lopez, P., Muñoz Sabater, J., Peubey, C., Radu, R., Richardson, D., Thépaut, J.-N., Vitart, F., Yang, X., Zsótér, E., and Zuo, H.: Operational global reanalysis: progress, future directions and synergies with NWP, Tech. Rep. 27, European Centre for Medium Range Weather Forecasts, https://doi.org/10.21957/tkic6g3wm, 2018.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J.,

Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J.-N.: The ERA5 Global Reanalysis, Quarterly Journal of the Royal Meteorological Society, –, *accepted for publication*, https://doi.org/10.1002/qj.3803, 2020.

Jiang, H., Yang, Y., Bai, Y., and Wang, H.: Evaluation of the total, direct, and diffuse solar radiations from the ERA5 reanalysis data in China, IEEE Geoscience and Remote Sensing Letters, p. *Early Access*, https://doi.org/10.1109/lgrs.2019.2916410, 2019.

Jiménez, 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 Atmospheres, 116, D02 102, https://doi.org/10.1029/2010JD014545, 2011.

Jiménez, C., Martens, B., Miralles, D. M., Fisher, J. B., Beck, H. E., and Fernández-Prieto, D.: Exploring the merging of the global land evaporation WACMOS-ET products based on local tower measurements, Hydrology and Earth System Sciences, 22, 4513–4533, https://doi.org/10.5194/hess-22-4513-2018, 2018.

Kobayashi, S., Ota, Y., Harada, Y., Ebita, A., Moriya, M., Onoda, H., Onogi, K., Kamahori, H., Kobayashi, C., Endo, H., Miyaoka, K., and Takahashi, K.: The JRA-55 reanalysis: general specifications and basic characteristics, Journal of the Meteorological Society of Japan. Ser. II, 93, 5–48, https://doi.org/10.2151/jmsj.2015-001, 2015.

Král, T.: Flux tower observations for the evaluation of land surface schemes: application to ERA-Interim, Tech. Rep. 11, European Centre for Medium Range Weather Forecasts, https://doi.org/–, 2011.

Kustas, W. and Daughtry, C.: Estimation of the soil heat flux/net radiation ratio from spectral data, Agricultural and Forest Meteorology, 49, 205–223, https://doi.org/10.1016/0168-1923(90)90033-3, 1990.

Liu, Y., Zhuang, Q., Pan, Z., Miralles, D., Tchebakova, N., Kicklighter, D., Chen, J., Sirin, A., He, Y., Zhou, G., and Melillo, J.: Response of evapotranspiration and water availability to the changing climate in Northern Eurasia, Climatic Change, 126, 413–427, https://doi.org/10.1007/s10584-014-1234-9, 2014.

Liu, Y. Y., de Jeu, R. A. M., McCabe, M. F., Evans, J. P., and van Dijk, A. I. J. M.: Global long-term passive microwave satellite-based retrievals of vegetation optical depth, Geophysical Research Letters, 38, L18 402, https://doi.org/10.1029/2011GL048684, 2011.

Liu, Y. Y., van Dijk, A. I. J. M., McCabe, M. F., Evans, J. P., and de Jeu, R. A. M.: Global vegetation biomass change (1988–2008) and attribution to environmental and human drivers, Global Ecology and Biogeography, 22, 692–705, https://doi.org/10.1111/geb.12024, 2013.

Martens, B., Miralles, D., Lievens, H., Fernández-Prieto, D., and Verhoest, N. E. C.: Improving terrestrial evaporation estimates over continental Australia through assimilation of SMOS soil moisture, International Journal of Applied Earth Observation and Geoinformation, 48, 146–162, https://doi.org/10.1016/j.jag.2015.09.012, 2016.

Martens, B., Miralles, D. G., Lievens, H., Van Der Schalie, R., De Jeu, R. A., Fernández-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, 10, 1903–1925, https://doi.org/10.5194/gmd-10-1903-2017, 2017.

Michel, D., Jiménez, 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, https://doi.org/10.5194/hess-20-803-2016, 2016.

Miralles, D., Gentine, P., Seneviratne, S., and Teuling, A.: Land-atmospheric feedbacks during droughts and heatwaves: state of the science and current challenges, Annals of the New York Academy of Science, 1436, 19–35, https://doi.org/10.1111/nyas.13912, 2018.

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, https://doi.org/10.5194/hess-15-453-2011, 2011.

Miralles, D. G., Teuling, A. J., Van Heerwaarden, C. C., and De Arellano, J. V. G.: Mega-heatwave temperatures due to combined soil desiccation and atmospheric heat accumulation, Nature Geoscience, 7, 345–349, https://doi.org/10.1038/ngeo2141, 2014.

Miralles, D. G., Nieto, R., McDowell, N. G., Dorigo, W. A., Verhoest, N. E. C., Liu, Y. Y., Teuling, A. J., Dolman, A. J., Good, S. P., and Gimeno, L.: Contribution of water-limited ecoregions to their own supply of rainfall, Environmental Research Letters, 11, 124 007, https://doi.org/10.1088/1748-9326/11/12/124007, 2016.

Miralles, D. G., Gash, J. H., Holmes, T. R. H., de Jeu, R. A. M., and Dolman, A. J.: Global canopy interception from satellite observations, Journal of Geophysical Research Atmospheres, 115, D16 122, https://doi.org/10.1029/2009JD013530, 2010.

Muñoz Sabater, J.: First ERA5-Land dataset to be released this spring, ECMWF newsletter 159, 2019.

Mueller, B., Hirschi, M., Jiménez, 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, https://doi.org/10.5194/hess-17-3707-2013, 2013.

Murphy, D. M. and Koop, T.: Review of the vapour pressures of ice and supercooled water for atmospheric applications, Quarterly Journal of the Royal Meteorological Society, 131, 1539–1565, https://doi.org/10.1256/qj.04.94, 2005.

Olauson, J.: ERA5: the new champion of wind power modelling?, Renewable Energy, 126, 322–331, https://doi.org/10.1016/j.renene.2018.03.056, 2018.

Priestley, C. and Taylor, R.: On the assessment of surface heat flux and evaporation using large-scale parameters, Monthly Weather Review, 100, 81–92, https://doi.org/10.1175/1520-0493(1972)100<0081:OTAOSH>2.3.CO;2, 1972.

Reichle, R., Koster, R., De Lannoy, G., Forman, B., Liu, Q., Mahanama, S., and Toure, A.: Assessment and enhancement of MERRA land surface hydrology estimates, Journal of Climate, 24, 6322–6338, https://doi.org/10.1175/JCLI-D-10-05033.1, 2011.

Reichle, R. H., Draper, C. S., Liu, Q., Girotto, M., Mahanama, S. P. P., Koster, R. D., and De Lannoy, G. J. M.: Assessment of MERRA-2 land surface hydrology estimates, Journal of Climate, 30, 2937–2960, https://doi.org/10.1175/JCLI-D-16-0720.1, 2017.

Santanello, J. and Friedl, M.: Diurnal covariation in soil heat flux and net radiation, Journal of Applied Meteorology, 42, 851–862, https://doi.org/10.1175/1520-0450(2003)042<0851:DCISHF>2.0.CO;2, 2003.

Scott, D.: On optimal and data-based histograms, Biometrika, 66, 605–610, https://doi.org/10.1093/biomet/66.3.605, 1979.

Seneviratne, S. I., Lüthi, D., Litschi, M., and Schär, C.: Land-atmosphere coupling and climate change in Europe, Nature, 443, 205–209, https://doi.org/10.1038/nature05095, 2006.

Seneviratne, S. I., Corti, T., Davin, E. L., Hirschi, M., Jaeger, E. B., Lehner, I., Orlowsky, B., and Teuling, A. J.: Investigating soil moisture-climate interactions in a changing climate: a review, Earth-Science Review, 99, 125–161, https://doi.org/10.1016/j.earscirev.2010.02.004, 2010.

Tarek, M., Brissette, F. P., and Arsenault, R.: Evaluation of the ERA5 reanalysis as a potential reference dataset for hydrological modelling over North America, Hydrology and Earth System Sciences, 24, 2527–2544, https://doi.org/10.5194/hess-24-2527-2020, 2020.

Taylor, C. M., de Jeu, R. A. M., Guichard, F., Harris, P. P., and Dorigo, W. A.: Afternoon rain more likely over drier soils, Nature, 489, 423–426, https://doi.org/10.1038/nature11377, 2012.

Tetzner, D. and Thomas, E.: A Validation of ERA5 reanalysis data in the Southern Antarctic Peninsula – Ellsworth land region, and its implications for ice core studies, Geosciences, 9, 289, https://doi.org/10.3390/geosciences9070289, 2019.

Teuling, A. J., Taylor, C. M., Meirink, J. F., Melsen, L. A., Miralles, D. G., van Heerwaarden, C. C., Vautard, R., Stegehuis, A. I., Nabuurs, G.-J., and de Arellano, J. V.-G.: Observational evidence for cloud cover enhancement over western European forests, Nature Communications, 8, 14 065, https://doi.org/10.1038/ncomms14065, 2017.

Teuling, A. J., Seneviratne, S. I., Stoeckli, R., Reichstein, M., Moors, E., Ciais, P., Luyssaert, S., van den Hurk, B., Ammann, C., Bernhofer, C., Dellwik, E., Gianelle, D., Gielen, B., Gruenwald, T., Klumpp, K., Montagnani, L., Moureaux, C., Sottocornola, M., and Wohlfahrt, G.: Contrasting response of European forest and grassland energy exchange to heatwaves, Nature Geoscience, 3, 722–727, https://doi.org/10.1038/NGEO950, 2010.

Tuanmu, M.-N. and Jetz, W.: A global 1-km consensus land-cover product for biodiversity and ecosystem modelling, Global Ecology and Biogeography, 23, 1031–1045, https://doi.org/10.1111/geb.12182, 2014.

Urraca, R., Huld, T., Gracia-Amillo, A., Martinez-de Pison, F. J., Kaspar, F., and Sanz-Garcia, A.: Evaluation of global horizontal irradiance estimates from ERA5 and COSMO-REA6 reanalyses using ground and satellite-based data, Solar Energy, 164, 339–354, https://doi.org/10.1016/j.solener.2018.02.059, 2018.

Valente, F., David, J. S., and Gash, J. H. C.: Modelling interception loss for two sparse eucalypt and pine forests in central Portugal using reformulated Rutter and Gash analytical models, Journal of Hydrology, 190, 141–162, https://doi.org/10.1016/S0022-1694(96)03066-1, 1997.

Vinukollu, R. K., Wood, E. F., Ferguson, C. R., and Fisher, J. B.: Global estimates of evapotranspiration for climate studies using multi-sensor remote sensing data: evaluation of three process-based approaches, Remote Sensing of Environment, 115, 801–823, https://doi.org/10.1016/j.rse.2010.11.006, 2011.

Wang, C., Graham, R. M., Wang, K., Gerland, S., and Granskog, M. A.: Comparison of ERA5 and ERA-Interim near-surface air temperature, snowfall and precipitation over Arctic sea ice: effects on sea ice thermodynamics and evolution, The Cryosphere, 13, 1661–1679, https://doi.org/10.5194/tc-13-1661-2019, 2019.

Wang, K. and Dickinson, R.: A review of global terrestrial evapotranspiration: observation, modeling, climatology, and climatic variability, Reviews of Geophysics, 50, RG2005, https://doi.org/10.1029/2011RG000373.1, 2012.

Wild, M., Folini, D., Hakuba, M. Z., Schär, C., Seneviratne, S. I., Kato, S., Rutan, D., Ammann, C., Wood, E. F., and König-Langlo, G.: The energy balance over land and oceans: an assessment based on direct observations and CMIP5 climate models, Climate Dynamics, 44, 3393–3429, https://doi.org/10.1007/s00382-014-2430-z, 2015.

Wouters, H., Petrova, I. Y., van Heerwaarden, C. C., Vilá-Guerau de Arellano, J., Teuling, A. J., Meulenberg, V., Santanello, J. A., and Miralles, D. G.: Atmospheric boundary layer dynamics from balloon soundings worldwide: CLASS4GL v1.0, Geoscientific Model Development, 12, 2139–2153, https://doi.org/10.5194/gmd-12-2139-2019, 2019.

Zhang, Y. and Cai, C.: Consistency evaluation of precipitable water vapor derived from ERA5, ERA-Interim, GNSS, and radiosondes over China, Radio Science, 54, RS006 789, https://doi.org/10.1029/2018RS006789, 2019.