Hydrol. Earth Syst. Sci. Discuss.,
https://doi.org/10.5194/hess-2020-595-AC1, 2021
© Author(s) 2021. CC BY 4.0 License.

Hydrology and Earth System Sciences
Discussions

Interactive comment on "Robust historical evapotranspiration trends across climate regimes" *by* Sanaa Hobeichi et al.

Sanaa Hobeichi et al.

s.hobeichi@unsw.edu.au

Received and published: 28 April 2021

We would like to thank the referees for their constructive comments on our manuscript. This document outlines our responses to their comments. We provide a track changed version of the manuscript to highlight the changes made to the manuscript and the supplementary material. In addition to the suggested changes by the two referees, we have further improved the analysis by introducing a parallel, complementary dataset version to DOLCE V2.1, DOLCE V3, that has fewer parent datasets than V2.1, reducing the number of temporal tiers and temporal discontinuities found in DOLCE V2.1, mostly over the tropics. DOLCE V2.1 remains a more optimal dataset in many senses as it minimises bias and maximises correlation with in-situ

observation, whereas V3 prioritises temporal continuity. Similar to DOLCE V2.1, the superiority of DOLCE V3 over its parents is demonstrated using an out-of-sample testing approach. DOLCE V3 is presented alongside DOLCE V2.1 throughout the manuscript and has not resulted in any new sections or qualitative change to the manuscript. The main change is in section '3.5 Changes in ET since 1980', in which DOLCE V3 was used instead of DOLCE V2.1 to carry out the analysis of trends. The new results show that trends in DOLCE V3 ET are mostly within the range of trends in available ET datasets, unlike DOLCE V2.1 whose temporal inconsistencies resulted in higher trends than the available datasets mostly over the wet ET regimes. We have amended related text, figures and tables accordingly. These updated results also help to address the concerns of the referees, as outlined below.

==========

Response to Referee 1, Jasper Denissen General Comments: ==========

Because of the efforts made to improve on DOLCE v1, a big part of the paper is about verifying DOLCE v2 against the parent datasets and in-situ observations. I think the title should reflect that. We agree with the referee in that the technical side of the paper which includes improving DOLCE, comparing it with its parents, and verifying it against in-situ observations constitute a big part of the paper. However, given that we are not publishing this work as a data paper, we chose a title that highlights the scientific side of the this work, which is the robust analysis of trends in ET. Furthermore, most of the figures that show the performance of DOLCE against in-situ observations and highlight its superiority over its parent datasets have now moved to the supplementary material as suggested by both referees, and as a result of this, 4 out of the 7 figures included in the main manuscript are now focused on the assessment of trends. The authors put a lot of effort on trying to find ways to improve DOLCE v2 in comparison to DOLCE v1 by i) weighting groups (Figure 1; right column) and ii) Bias correction strategies (Figure 2; Supplementary Figure 3) as described in 2.2.4, 2.2.5 and 3.1. Despite the authors efforts to significantly improve the ET estimates, I think that the added complexity does not justify the little improvement gained. I would suggest moving sections 2.2.4,

2.2.5 and 3.1 and corresponding figures to the Supplementary materials. In the main text, the authors can shortly motivate the weighting group and bias correction strategy by referring to the Supplementary materials. This would make the derivation of DOLVE v2 ET more straight-forward and benefit the readability of the manuscript. We agree with the referee that the clustering methods added little improvement to the weighting, however a little improvement is better than nothing. Technically, grouped weighting requires aggregating the time and/or space domains prior to applying the weighting technique and adds a small amount of work and a few additional seconds of processing time. It is therefore worth investigating the efficacy of grouping approaches, and choosing the methods that adds the most improvement to the weighting even if the improvement is marginal. Furthermore, as detailed in the text, most of the weighting strategies have been previously suggested as ways to improve merging but have not been tested. Therefore, testing them here provides valuable information to the science community. To address the referee's concern and increase the readability of the manuscript, we have now moved sections '2.2.5 Bias correction strategies' and '3.1 Selection of a grouping strategy' and associated figures to the supplementary material.
==========

Specific comments: ========== Throughout the paper, please properly introduce i) table contents, ii) figure axes labels and color codes and iii) statistics of box-plots. We thank the referee for his comment. We have now explained the figures and the tables further throughout the text and in the captions. lines 20: I found the notion that these climatology clusters / climate regimes are able to summarize or even replace the Köppen-Geiger climate regimes quite interesting. That would be worth a mention in the abstract, also putting this sentence into context. We thank the referee for his suggestion, we have now mentioned the agreement of these classes with the Köppen-Geiger climate regimes in the abstract .... The new clusters include three wet and three dry regimes and provide an approximation of Köppen-Geiger climate classes. lines 23-25: "We find that despite robust . . . ET clusters". This is the only time this is mentioned in the entire manuscript. I don't see the relevance

of it for the abstract. Good point. We have now shown this finding in section 3.5.3 Global annual trends across the ET regimes ... Our results indicate that decreasing ET trends observed in some regions oppose the consistent positive trends across the majority of ET clusters. lines 82-84: FLUXCOM also belongs in this summation of gridded datasets that successfully exploit in-situ measurements. Here we are listing the studies that applied fusion techniques attempting to match a global dataset that is deemed more reliable than the original datasets. FLUXCOM was listed earlier in the introduction with the machine-learning based datasets: .... techniques including machine-learning algorithms (Jung et al. 2010; Hamed Alemohammad et al. 2017; Jung et al. 2019), typically incorporating a range of remote sensing inputs. lines 198-200: I assume that the 'very large spatiotemporal domains' are equal to the spatial and temporal resolutions of the ET data sets? Do the authors mean that through time and varying wind directions, you might actually get closer to the grid cell mean than looking at individual days? We thank the referee for his comment. 'Very large spatiotemporal domains' means over many sites and time steps. This paragraph is trying to say that weights are derived by assessing the agreement between flux tower measurements and the value of underlying grid cells over many locations and time steps. This however assumes that the point scale of the flux towers can represent the grid scale. We don't expect this representativeness to be true at each site, however the ensemble of flux tower observations as a whole do represent the underlying grid cells. This has been thoroughly tested and validated in previous work. We have now made the paragraph clearer: First, weights for each product are constructed over very large spatiotemporal domains, i.e. more than 13000 space-time records as described below, so that the (assumed stochastic) biases of individual sites relative to grid cell values are unlikely to influence weights over a large sample. In fact representativeness of point-scale measurement for the grid scale does exist across all the flux tower sites as a whole, this has been verified by (Hobeichi et al. 2018). lines 223-230: Just out of interest: What is the total amount of days initially available from all sites? After the filtering based on data availability, how many days are left? The original sites data

was a mix of half-hourly, daily and monthly data. The majority of daily data come from Ameriflux sites. The raw Ameriflux data consisted of 147 sites with daily data and a total of 191,583 daily records. After quality control and filtering, the number of sites and daily records dropped to 56 and 81,142 respectively. line 241: Could you elaborate on these conditions? As I'm not a flux tower measurements expert: Are these 20 and 30 W m-2 thresholds usually applied? Is there a paper where this methodology is also applied? Good point, our response to this comment is now included in the text A study by Paca et al. (2019) examined the changes to flux tower LE by three means of corrections, and found that these on average differ by around 20 Wm-2 from one another. On this basis, we expect that typically, the correction of flux tower LE should not exceed 20 Wm-2, unless errors in other components of the budgets are propagating in the corrected ET. The rule for correcting small fluxes and the condition in which each rule is applied (i.e. LE= 30 Wm-2) are in part subjective and in another part based on a case by case assessment of changes induced to ET by the correction techniques, and achieve a reasonable trade-off between data quality and availability. lines 246-249: How do you justify using LE values without any correction – any value is better than nothing? Did you verify the differences between i) only LE data with correction and ii) LE data with and without correction? Are there any biases there? All ET measurements are prone to systematic errors, here we are using the physical constraints of the energy balance to minimize these errors. Constraining ET this way is a 'plus' rather than a 'must'. ET measurements, despite systematic errors in them, provide the most reliable information on ET and can still be used for ground-truthing gridded ET estimates. There is certainly a bias between i) and ii), however the majority of the sites, this bias is small relative to the values in gridded estimates. lines 250-255: Why not just take an average of the different towers within one grid cell, weighted by fractional cover of biome within that grid cell? I would assume that you want to have the possibility of retaining as much data as possible, as there could be data gaps in the data records in the tower measurements which could be filled with values from towers in the vicinity. And an additional question: Do you somehow account for

varying footprints for the in-situ ET measurements? Depending on the location of the flux tower within the grid cell, what the tower measures could be from a neighboring grid cell. Weighting ET measurements from two neighbouring towers located in two different biomes, with one dominant and another less dominant is expected to provide a better ET than only taking ET from the dominant biome. However, in all these cases, flux towers have different years of coverage. Therefore, considering data from both sites will retain more data but at the same time will lead to temporal inconsistency in the timeseries of the combined ET. No, we haven't accounted for varying footprints for the in-situ measurements. However, we have visually assessed the position of each flux tower in the grid cell using ArcGIS, and we have found that of the 260 sites used in this study, only two sites are close to the edge of the grid which means that their footprint might extend to that neighbouring grid cell. However, at both sites, the land cover and the climate within the expected footprint of the tower (< 2000 m) across the underlying and neighbouring grid cells is the same. lines 272-275: Are w_k are the weights, which sum to 1, based on the error correlation in Supplementary Figure 2? If so, please state that here for clarity. We thank the referee for his suggestion. We have now added a reference to Figure S2 in the text The analytical solution to this problem accounts for both the performance differences between the parent datasets and their error covariance (Fig. S2), a proxy for dependence. lines 296-301: "the discrepancy between DOLCE and actual ET at any spatial scale greater than that of a tower footprint should be less than this uncertainty estimate" I am slightly confused here: This would only be true for spatial scales greater than the tower footprint and smaller than 0.25x0.25 degree grid cell res- olution (which is used in the study), right? If the spatial scale would be greater than 0.25x0.25 degrees, the discrepancy should be even larger? The referee is right. We have now made this clear in the text ... the discrepancy between DOLCE and actual ET at any spatial scale greater than that of a tower footprint and smaller than that of DOLCE should be less than this uncertainty estimate lines 351-353: This might be a naïve question, but is that not just because of the small number of flux towers on the SH? I could imagine that due to

the higher availability of in-situ measurements and abundance of other measurements net- works ET estimates are much more well constrained in the NH than the SH. This is true to a large degree for the data driven approaches such as machine learning ET estimates, which will have their reliability degraded in areas with less observations. However, we cannot assert that the lower availability of in-situ measurements in the Southern Hemisphere is driving disagreement among modelled and remote sensing ET estimates given that these typically do not rely on flux measurements. 357-359: I was confused here, as I assumed the authors were referring to boreal summer. Please clarify. We thank the referee for spotting this. Incorrect months were originally attributed to summer-fall and winter-spring seasons in the Northern and Southern hemispheres. We have now corrected the text by listing the correct months in each season We consider two combined seasons i.e. summer-fall and winter-spring. In the summer-fall season, we constrain the weighting with (1) monthly observations from sites located in the Northern hemisphere during the period June–November, and (2) monthly observations from sites located in the Southern hemispheres during the December–May. The remaining observational data is used to constrain the weighting during the winter-spring combined season. lines 380-282: I do not understand what is meant with 'extrapolating the bias field'. Please explain more clearly. The Bias field is the ET bias values described in lines 378 – 380. We have spatially extrapolated the ET bias values from the grid cells containing the sites to the entire global land. We have now made this clearer in the text. We then assign those ET bias values, or bias field, to the grid cells containing the sites. Finally, using the bias values at these grid cells, we extrapolate the bias field spatially to the entire global land domain within the tier using several different extrapolation strategies,... Figure 3: Does zonal ET follow a Gaussian distribution? If not, it would make more sense to define the grey ribbon as the interquartile range instead of the standard deviation. Also, I would suggest making a mask of the grid cells where all of the parent data sets have values, so that a comparison is fairer. The figure without the mask could then be moved to the supplementary material. We haven't examined the distribution of zonal ET,

and the uncertainty standard deviation term is not describing the latitudinal spread of ET around the mean (i.e. zonal standard deviation), but rather it shows the range of DOLCE ET values bounded by its uncertainty, i.e. DOLCE $\pm$ uncertainty. The uncertainty of DOLCE is explained in Section 2.2.2, and is described as the 'standard deviation of uncertainty'. We have now omitted 'standard deviation' from the caption to ensure that the reader does not misinterpret the grey ribbon and clarified in the text that the grey ribbon is 'DOLCE $\pm$ uncertainty'. We have also replaced this plot with a new version that applies the same spatial mask to all datasets. In the new figure, we have now added a similar plot for DOLCE V3 and its parents. lines 500-501: Mentioning the seasonal cycle of DOLCE v2 ET but not elaborating on it feels out of place. Either elaborate on differences between seasonal cycles between DOLCE v2 ET and others or remove. We thank the referee for his suggestions. We have now removed the plot of seasonal climatology lines 525-528: Please clarify this sentence; I found it confusing as written. We thank the referee for his comment. We have now clarified the sentence. The uncertainty estimate of DOLCE, however, is firmly grounded in the discrepancy between the gridded DOLCE product and in-situ tower data. The variance of this discrepancy is used to recalibrate the variance of the parent datasets, which are then used to estimate uncertainty, allowing spatiotemporally varying uncertainty estimate that is both consistent with the discrepancy between DOLCE and surface observations while at the same time being spatially and temporally complete. This process is detailed by Hobeichi et al (2018). lines 540-542: Either put into context by comparing with all the other literature references or remove. These two sentences seem lost. We thank the referee for his suggestion. We have now removed lines 540 - 542 from the text lines 546-550: In the first sentence the authors state the RMSE is not computed because the means between DOLCE and sites are not equal. In the sentence after you explain that all data has been normalized and therefore all have a zero mean. So, by normalizing you could in principle calculate the RMSE? We used the Taylor diagram to show how DOLCE performs at all sites, rather than a single site. To do this it was necessary to normalise the data because there will be different values of

the observed variable across different sites. Since the data is now normalised RMSE is not helpful here. To avoid any confusion, we have now removed this sentence from the paragraph. lines 558-560: Would the signal of land cover be clearer when the authors would aggregate the land cover types to short/tall vegetation? Next to that, in Supplementary Figure 6, the color legend is blocking some of the extreme values. We thank the referee for suggesting this. We have now combined 6 classes of broadleaved and needle leaved tree covers in one group, but we still could not find any clear signal. We have made the fill color of the legend transparent to make sure no values are covered. lines 571-572: Is there a specific conclusion drawn from these figures? Otherwise mentioning them seems unnecessary to me. We thank the referee for his suggestion. We have now removed these plots. Have the authors also looked at ET trends from flux tower measurements? As trends across all KS-clustering defined climate regimes are positive, the flux tower observations could corroborate that if they are also generally positive, right? Good point. Flux tower measurements have only started around 2002, and the longest observational records we had at any site was only 17 years. Therefore, there is no in-situ observations that cover a long enough period to examine trends in annual ET from observations. This was already acknowledged in the text in section 3.5.1 : Unfortunately, given the absence of adequate in-situ observations that cover a long enough period to establish trends analysis, it is difficult to validate the identified trends directly. line 678-680: I don't know how the fact that the global ET trends are different than the other ET products reflects usefulness; the fact that the DOLCE trends are different does not necessarily mean they are correct. However, it would be really interesting to see whether flux tower observations find similar long-term ET trends. We agree with the referee in that trends in DOLCE V2.1 are much stronger than those in other datasets. After some additional investigation we have made significant changes to this section, as explained below. However we do note that Table 5 shows that the global ET trends in the other datasets show differences in sign, magnitude and statistical significance of the trends across all ET regimes. We have now introduced DOLCE V3, a complementary dataset to

DOLCE V2.1, which we now use to carry out the analysis of trends instead of DOLCE V2.1. The main difference between the two datasets is the number of contributing parent datasets. The focus for DOLCE V3 was reducing the number of temporal tiers to reduce temporal discontinuities in DOLCE V2.1 (these were revealed in a separate analysis not related to this manuscript), mostly over the tropics. We believe these discontinuities and inhomogeneity lead to misrepresentative trends in some cases. We present V3 as a parallel version (rather than replacement) as V2.1 still remains a better performing data set in many of the out of sample tests. We have now repeated the analysis of trends in Section 3.5 and have shown new trends results incorporating DOLCE V3 (instead of DOLCE V2.1). The new results show that trends in DOLCE V3 agree with some products more than the others, and its trends' slopes are within the range of slopes of trends in the available products across all the ET regimes. This is now shown in the text and in the updated Table 5 We repeat the same analysis for all the participating parent datasets that span at least 30 years. Sen's slope of the trends and their confidence interval (computed at the 95% confidence level) are presented in Table 5. As noted earlier, trends' behaviour is deemed inconclusive when the CI encompasses negative and positive values. These are presented with regular (as opposed to bold) typeface and are exhibited by FLUXCOM-MET in all regimes except the driest. ERA5-land shows downward trends in the 'M.H.ET, M.variability' and 'H.ET, L.variability' regimes. Both GLEAM 3.5A and PLSH show upward ET trends in all regimes, with the exception of GLEAM which shows no reliable trends in the driest and wettest ET regimes. Differences exist in the magnitude of trends across the majority the products and the regimes. As in DOCLE V3, the strongest trends in GLEAM 3.5A occur in the 'M.H.ET, M.variability' regime at a rate 0.5 mm ãĂŰyearãĂů^(-1). Finally, the slopes of DOLCE V3 trends are within the range of slopes of trends in available ET products. Also, none of the available datasets incorporate the same degree of observational constraint in either their mean field or uncertainty estimates, which makes us believe that trends exhibited by DOLCE V3 are more reliable than those observed in other datasets. As we mentioned in our response to the previous

comment above, there are not enough in-situ observations that cover a long enough period to examine trends in annual ET directly from observations.

Technical corrections

line 21: ‘at each location’. Do you mean globally?

Yes, we have now made it clear in the text by adding ‘on land’. The sentence now reads:

. . . we derive novel ET climatology clusters for the land surface, based on the magnitude and variability of ET at each location on land.

line 42: remove the comma after ‘approaches’ We thank the referee for spotting this. We have now removed the coma after ‘approaches’. line 113: replace ‘trends (5) behavioural’ with ‘trends and (5) behavioural’ Thank you for suggesting this. We have now made the change in the text. lines 236: to avoid confusion, maybe rephrase “latent heat measurements are used directly” to “latent heat measurements are used without any corrections”. We thank the referee for his suggestion. We have now replaced ‘directly’ with ‘without any corrections’. line 435: Replace ‘Fig. S3’ with ‘Fig S4.’ We thank the referee for spotting this. We have now made the change in the text. line 590: replace ‘intensified’ with’ increased’ We thank the referee for his suggestion. We have now made the change. line 624: replace ‘modifed’ with ‘modified’ We thank the referee for spotting this. We have now made the correction. line 743: replace ‘Figure2’ with ‘Figure 2’ We thank the referee for spotting this. We have now added a space after ‘Figure’.

Please also note the supplement to this comment:
https://hess.copernicus.org/preprints/hess-2020-595/hess-2020-595-AC1-supplement.pdf

---

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2020-595, 2020.

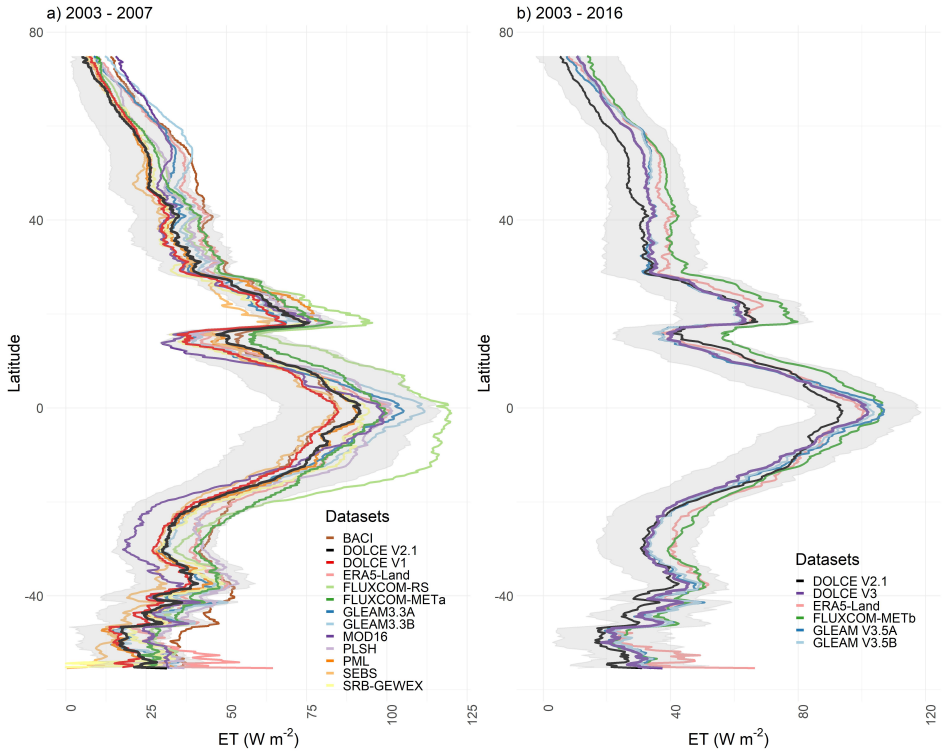
a) 2003 - 2007
b) 2003 - 2016
Latitude
ET (W m-2)
Datasets
BACI
DOLCE V2.1
DOLCE V1
ERA5-Land
FLUXCOM-RS
FLUXCOM-METa
GLEAM3.3A
GLEAM3.3B
MOD16
PLSH
PML
SEBS
SRB-GEWEX
DOLCE V3
FLUXCOM-METb
GLEAM V3.5A
GLEAM V3.5B

Fig. 1.

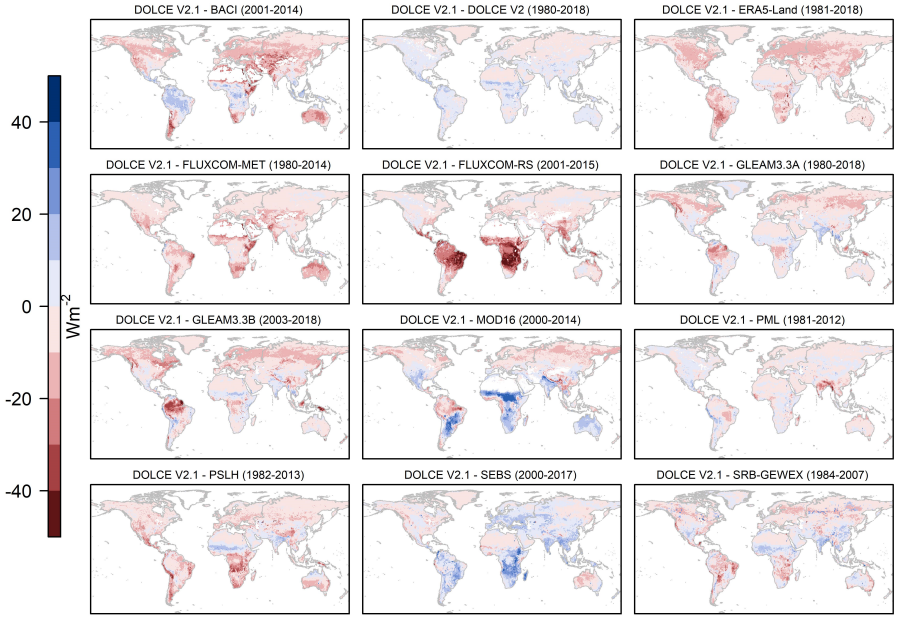
DOLCE V2.1 - BACI (2001-2014)
DOLCE V2.1 - DOLCE V2 (1980-2018)
DOLCE V2.1 - ERA5-Land (1981-2018)
DOLCE V2.1 - FLUXCOM-MET (1980-2014)
DOLCE V2.1 - FLUXCOM-RS (2001-2015)
DOLCE V2.1 - GLEAM3.3A (1980-2018)
DOLCE V2.1 - GLEAM3.3B (2003-2018)
DOLCE V2.1 - MOD16 (2000-2014)
DOLCE V2.1 - PML (1981-2012)
DOLCE V2.1 - PSLH (1982-2013)
DOLCE V2.1 - SEBS (2000-2017)
DOLCE V2.1 - SRB-GEWEX (1984-2007)
Wm-2

Fig. 2.

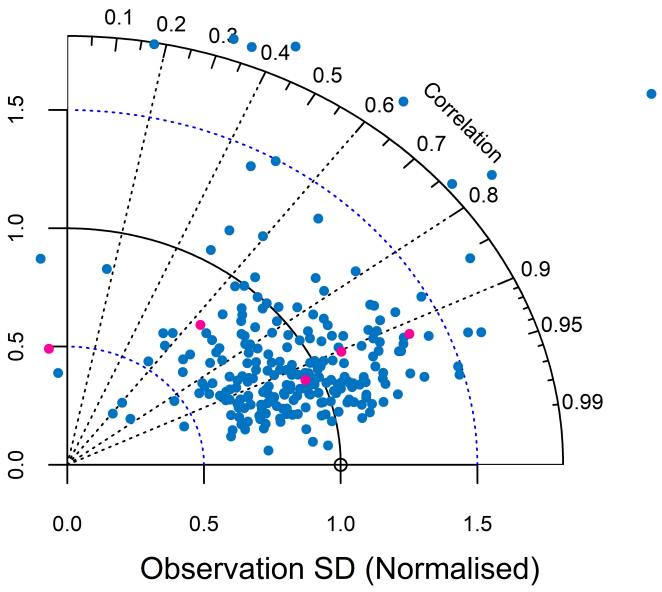
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
0.95
0.99
Correlation
0.0
0.5
1.0
1.5
0.0
0.5
1.0
1.5
Observation SD (Normalised)

Fig. 3.

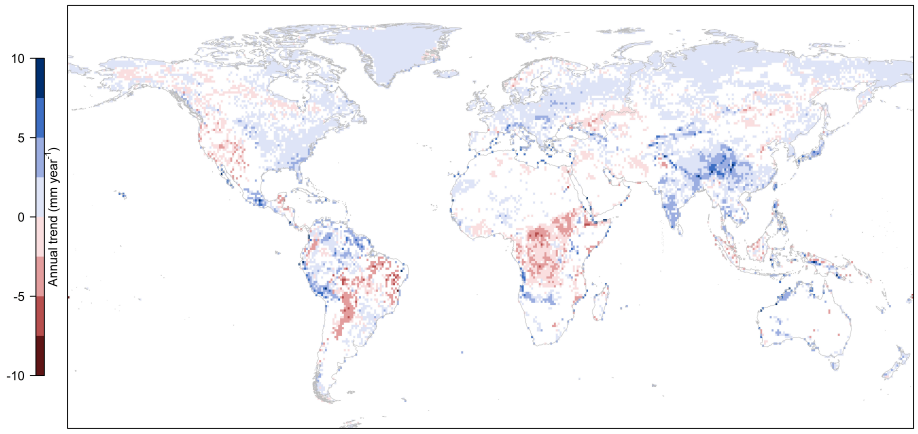
10
5
0
-5
-10
Annual trend (mm year⁻¹)

Fig. 4.

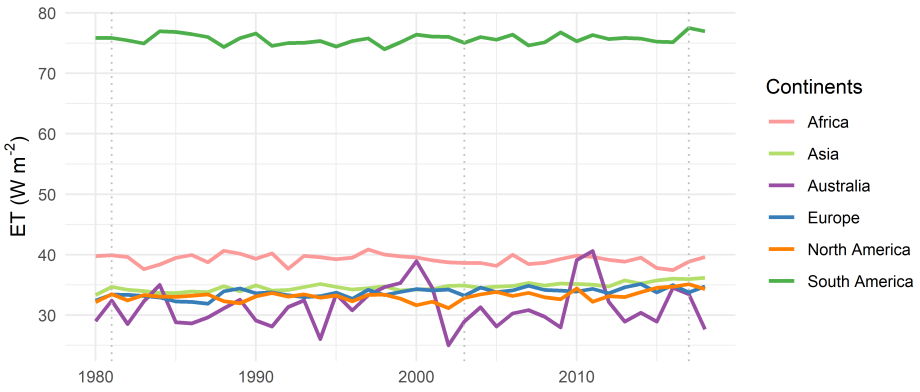
80
70
60
50
40
30
ET (W m-2)
1980
1990
2000
2010
Continents
Africa
Asia
Australia
Europe
North America
South America

Fig. 5.

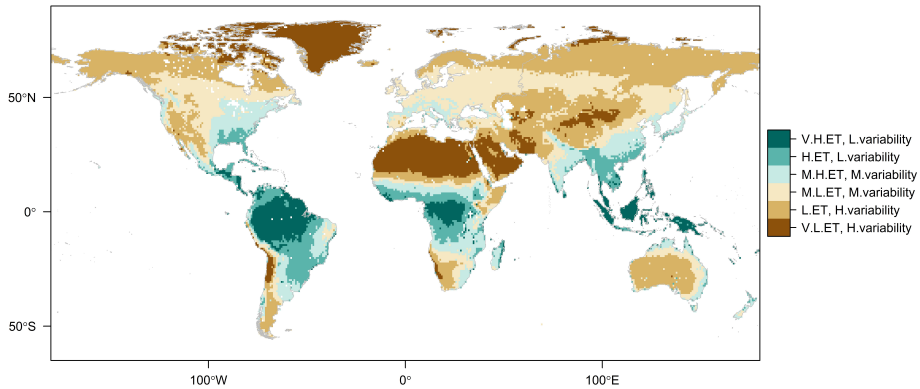
50°N
0°
50°S
100°W
0°
100°E
V.H.ET, L.variability
H.ET, L.variability
M.H.ET, M.variability
M.L.ET, M.variability
L.ET, H.variability
V.L.ET, H.variability

Fig. 6.

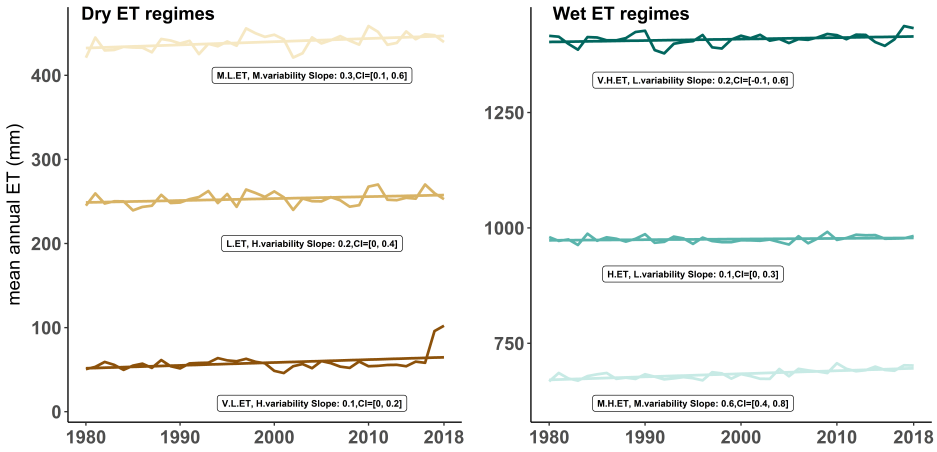
Dry ET regimes
Wet ET regimes
mean annual ET (mm)
M.L.ET, M.variability Slope: 0.3,CI=[0.1, 0.6]
L.ET, H.variability Slope: 0.2,CI=[0, 0.4]
V.L.ET, H.variability Slope: 0.1,CI=[0, 0.2]
V.H.ET, L.variability Slope: 0.2,CI=[-0.1, 0.6]
H.ET, L.variability Slope: 0.1,CI=[0, 0.3]
M.H.ET, M.variability Slope: 0.6,CI=[0.4, 0.8]
0
100
200
300
400
750
1000
1250
1980
1990
2000
2010
2018

Fig. 7.