Robust historical evapotranspiration trends across climate regimes

Sanaa Hobeichi[1,2], Gab Abramowitz[1,2], and Jason Evans[1,2]

[1] Climate Change Research Centre, UNSW Sydney, NSW 2052, Australia.

[2] ARC Centre of Excellence for Climate Extremes, UNSW Sydney, NSW 2052, Australia.

Correspondence: Sanaa Hobeichi (s.hobeichi@unsw.edu.au)

Abstract

Evapotranspiration (ET) links the hydrological, energy, and carbon cycle on the land surface. Quantifying ET and its spatiotemporal changes is also key to understanding climate extremes such as droughts, heatwaves and flooding. Regional ET estimates require reliable observationally-based gridded ET datasets, and while many have been developed using physically-based, empirically-based and hybrid techniques, their efficacy, and particularly the efficacy of their uncertainty estimates, is difficult to verify. In this work, we extend the methodology used in Hobeichi et al. (2018) to derive a two new versions of the Derived Optimal Linear Combination Evapotranspiration (DOLCE) product, with observationally constrained spatiotemporally varying uncertainty estimates, higher spatial resolution, more constituent products and extended temporal reach coverage (1980-2018). After successful evaluation ofdemonstrating the efficacy of these uncertainty estimates out-of-sample, we derive novel ET climatology clusters for the land surface, based on the magnitude and variability of ET at each location on land. The new clusters include three wet and three dry regimes and provide an approximation of Köppen-Geiger climate classes. The verified uncertainty estimates and extended time period then allow us to examine the robustness of historical trends spatially and in each of these six ET climatology clusters. We find that despite robust decreasing ET trends in some regions, these do not correlate with behavioural ET clusters. Each cluster, and the vast majority of the Earth's surface, show clear robust increases in ET over the recent historical period. The new datasets DOLCE V2.1 and DOLCE V3 can be used for benchmarking global ET estimates and for examining ET trends respectively.

1. Introduction

Understanding the spatiotemporal variability of evapotranspiration (ET) is a critical part of understanding the processes that lead to high impact weather phenomena, such as droughts (Han et al., 2018; Montano et al., 2015; Sheffield et al., 2012; Teuling et al., 2013), heatwaves (Teuling, 2018; Ukkola et al., 2018) and flooding (Dawdy et al., 1972; Sharma et al., 2018). Several global gridded ET datasets have been developed, using physical schemes with different scopes (e.g. addressing key questions in ecology, hydrology, or other disciplines), and complexity (see Fisher and Koven, 2020), and empirical

37   techniques including machine-learning algorithms ~~(Hamed Alemohammad et al. 2017; Jung et al. 2010,~~
38   ~~2019)~~, typically incorporating a range of remote sensing inputs (Hamed Alemohammad et al., 2017; Jung
39   et al., 2010, 2019). Recently, ET datasets derived with a hybrid approach have been recognised for their
40   potential to outperform single source datasets in reducing bias against tower-based eddy-covariance ET
41   measurements (Ershadi et al., 2014; Feng et al., 2016; Hobeichi et al., 2018; Jiménez et al., 2018;
42   McCabe et al., 2016).

43

44   While most observational products are global (or near global) in their spatial extent, and typically
45   available with a monthly time step, different products are constrained by very different types of
46   observations, and vary significantly in their treatment of uncertainty. As detailed below when describing
47   the datasets we use here, 'physically-based' approaches, use equations that represent different
48   physical, chemical, and biological processes and incorporate satellite-based atmospheric forcing, and
49   parameterization of land surface characteristics, while 'empirical' approaches integrate ground-based
50   measurements of ET together with satellite data and ground-based measurements of vegetation
51   characteristics and land surface parameters. These differences result in a diverse group of products and
52   estimates, but it is their approach to deriving uncertainty estimates that is arguably more important.

53

54   Very few datasets provide uncertainty estimates associated with the ET flux, these include datasets
55   described in Bodesheim et al. (2018) and Jung et al. (2019). In Bodesheim et al. (2018), monthly
56   uncertainty estimates are computed from the standard deviation of the half-hourly ET values that were
57   used to derive monthly ET averages. Jung et al. (2019) provide an ensemble of global ET estimates,
58   deviations from the ensemble median are used to derive ET uncertainties. In both cases, uncertainties
59   do not reflect the actual deviation from the measured ET at site locations. Without well calibrated
60   uncertainty estimates we are unable to tell whether an identified property of any given data set, such as
61   a trend or a proportion of the surface energy or water budget, is robust, rather than a result of bias or
62   stochastic uncertainty.

63

64   ET trends computed from different approaches (i.e. physical and empirical) show general agreement at
65   the global scale, and indicate that ET has increased since early 1980s (Miralles et al., 2014; Pan et al.,
66   2020; Zhang et al., 2016). However, different ET products exhibit considerable disparities in regional and
67   continental ET trends. For instance, Miralles et al. (2014) detected upward ET trends in GLEAM (Global
68   Land Evaporation Amsterdam Model; Miralles et al. 2011) in the northern latitudes caused by vegetation
69   greening. In water limited regions, they found that ET is characterised by a multidecadal variability that
70   follows ENSO dynamics, mainly in eastern and central Australia, southern Africa and eastern South
71   America. In comparison, ET trends estimated from the observation-driven Penman-Monteith-Leuning
72   (PML; Zhang et al. 2016) model show increasing ET since 1980 in the northern latitudes, arid regions in
73   northern Africa, and northern and eastern Amazon. On the other hand, PML exhibits negative trends in
74   southern South America and western United States. More recently, Pan et al. (2020) found that ET
75   trends exhibited during 1982-2011 by a range of empirical and physically-based estimates disagree in
76   the direction of trend in the Amazon basin and many arid and semi-arid regions. Without incorporating
77   uncertainties in ET estimates in the analysis of trends, it becomes difficult to assess the reliability of the
78   established trends.

79

80 The gridded ET product derivation technique implemented by Hobeichi et al. (2018) offers the potential
81 for robust out-of-sample testing of its uncertainty estimates, as well as several other advantages over
82 other techniques. Like other merging approaches, it offers the potential to minimise the eccentricities or
83 biases of any one product, by averaging them (in this case using weights). However, unlike several other
84 merging techniques (Mueller et al., 2013; Paca et al., 2019; Rodell et al., 2015; Stephens et al., 2012) it
85 accounts for performance differences between parent estimates using in-situ data as the observational
86 constraint, rather than assigning weights based on the ability to match another gridded dataset that is
87 deemed more reliable, or the ensemble mean of a selection of datasets (Munier et al., 2014; Sahoo et
88 al., 2011; Wan et al., 2015; Zhang et al., 2018). The efficacy of using in-situ measurements for
89 constraining much larger scale gridded estimates has also been shown explicitly (Hobeichi et al., 2018,
90 2020b). Next, most available merging techniques do not account for dependence between parent
91 estimates, where redundant information in different parent products is likely to bias the hybrid estimate
92 (Abramowitz et al., 2019; Herger et al., 2018). Finally, and perhaps most important for this work, the
93 technique calculates global spatially and temporally varying uncertainty estimates that are
94 observationally-based, in that they are based on the discrepancy between the hybrid ET estimate and in-
95 situ data. Aside from being more defensible than simply taking the spread of the parent products
96 around their mean (e.g. Pan et al., 2012, Zhang et al., 2018), this approach also allows for out-of-sample
97 testing, by leaving some sites out of the derivation of the hybrid product and its uncertainty, and then
98 using them to test its accuracy.
99
100 Despite these advantages, out-of-sample testing of uncertainty estimates was not explored by Hobeichi
101 et al (2018), and the short temporal availability of the DOLCE product (2000 – 2009) limited its
102 application, particularly in examining historical trends. While different subsets of parent products were
103 used over different regions to expand the spatial coverage of DOLCE, the possibility of different product
104 subsets in different time periods to extend its temporal reach was not explored. Additionally, since the
105 development of DOLCE, four of its six parent datasets (Jung et al., 2010; Martens et al., 2016; Miralles et
106 al., 2011; Mu et al., 2011; Zhang et al., 2016) have been improved and several new global ET datasets
107 have been developed (Balsamo et al., 2015; Bodesheim et al., 2018; Jung et al., 2019). Most of these are
108 available at a higher spatial resolution than the original 0.5° in DOLCE and cover different subsets of the
109 period 1980 – 2018, with at least two available ~~for~~ every year during this period (Table1).
110
111 In this paper we amend these shortcomings and explore some of the insights that the new version~~s~~ of
112 DOLCE offer~~s~~, in particular focusing on the temporal trends in ET in different regions, and the
113 assessment of robustness of trends that well calibrated uncertainty estimates afford. Roughly in order,
114 we detail below: (1) how we update the DOLCE product with new parent datasets and~~,~~ extend its~~ed~~
115 temporal coverage ~~and higher spatial resolution~~; (2) how the improved product~~s~~ compare~~s~~ to their~~its~~
116 previous version and other existing ET estimates from the literature; (3) the efficacy of uncertainty
117 estimates, in particular whether or not they are overconfident; (4) an exploration of historical trends in
118 ET using the extended temporal coverage, and how the uncertainty estimates allow us to examine the
119 robustness of these trends; and (5) behavioural ET clusters that describe ET based climate regimes, as a
120 mean to understand the spatial distribution of trends we find.
121
122

2. Data and Methods

To derive ~~a~~ two new version~~s~~ of DOLCE, one suitable for benchmarking ET dataset and another for trends analysis, we combine 11 and 4 available global gridded ET datasets respectively using the same merging technique as in DOLCE V1. This technique derives a linear combination of the participating ET datasets based on their ability to match in-situ observations while also accounting for their error dependency. While we acknowledge the obvious spatial mismatch between gridded and in-situ data, we refer readers to Hobeichi et al (2018) where it was shown that in-situ observations do contain useful information about grid scale fluxes, using out-of-sample testing in a similar framework to the one we present here.

Our aim is to increase the time coverage and spatial resolution of DOLCE V1, as well as examine strategies to improve the effectiveness of the weighting strategy. Below we detail newly available global datasets that allow us to derive DOLCE V2 and DOLCE V3 at 0.25° spatial resolution, and an improved collection of in-situ constraining data. We then briefly revisit the weighting and uncertainty estimation approach, before describing our tiering approach to extending the temporal reach of DOLCE V2 and DOLCE V3. Finally, we examine alternative clustering and bias-correction approaches to improve the out-of-sample performance of the weighting technique.

Throughout the paper, we use the two terms evapotranspiration (ET) and latent heat (LE) interchangeably, and the unit $W\ m^{-2}$ for heat fluxes and $mm\ year^{-1}$ for the water flux equivalent. For reference: $1\ W\ m^{-2} = 12.86\ mm\ year^{-1}$. As above, we refer to the product from Hobeichi et al (2018) as DOLCE V1 and the new product~~s~~ we are deriving as DOLCE V2 or DOLCE V2.1 and DOLCE V3~~.~~

2.1 Data

2.1.1 Global ET datasets:

DOLCE V1 was derived from 6 global ET datasets: MPIBGC (Jung et al., 2010), GLEAM v2a, GLEAM v2b (Miralles et al., 2011), GLEAM v3a (Martens et al., 2016, 2017), MOD16 (Mu et al., 2011) and PML (Zhang et al., 2016). In DOLCE V2, we keep both MOD16 and PML datasets, substitute the GLEAM products with their improved ~~and latest~~ versions ~~(i.e.~~ GLEAM3.3A and GLEAM3.3B~~;~~ (Martens et al., 2016, 2017), and replace MPIBGC with newly developed empirical ET datasets from the Max Planck Institute for Biogeochemistry: BACI (Bodesheim et al., 2018) and two ET estimates from the FLUXCOM project (Jung et al., 2019)~~s~~. Additionally, we incorporate a recently published dataset ERA5-Land (Muñoz Sabater, 2019) and three newly available ET datasets PLSH (Zhang et al., 2015), SEBS (Chen et al., 2019; Su, 2002) and SRB-GEWEX (Vinukollu et al., 2011). In comparison, DOLCE V3 was derived from 4 global ET datasets. These ~~include~~are: ERA5-Land, an ET dataset from the FLUXOM project, and the two latest versions of the GLEAM products, GLEAM V3.5A and GLEAM V3.5B ~~which became available at the time where DOLCE V3 was being developed.~~ We provide a brief description of these datasets below, with URLs and download dates shown in supplementary Table S2.

163   Biosphere Atmosphere Change Index (BACI; Bodesheim et al., 2018): The dataset is derived by upscaling
164   diurnal cycles of ET and other land-Atmosphere fluxes from a large set of FLUXNET sites based on a
165   random forest regression framework. It uses seasonal vegetation variables and indices from MODIS
166   satellites, and meteorological data either measured at the flux tower sites or retrieved from the ERA-
167   Interim data.

168   ERA5-Land (Muñoz Sabater, 2019)(Balsamo et al. 2015): A global land surface reanalysis dataset that has
169   been developed by rerunning the land component of the ECMWF ERA5 climate reanalysis with a series of
170   improvements (mainly higher temporal frequency and spatial resolution) that makes it more reliable for
171   land applications. ERA5-Land is produced under a single simulation that uses adjusted atmospheric inputs
172   from ERA5 atmospheric variables without being coupled to the atmospheric module of ERA5.

173   FLUXCOM (Jung et al., 2019): An empirical upscaling of observations from 224 flux tower sites using
174   machine learning methods. The full FLUXCOM product includes 63 global ET datasets that have been
175   produced using two different setups, a remote sensing (RS) setup and a remote sensing + meteorological
176   (MET) setup. The development of the global datasets incorporates 9 machine learning techniques, 4 global
177   meteorological datasets (used only with the MET setup), 3 correction methods for energy imbalance at
178   the flux tower sites and MODIS remote sensing input. In DOLCE V2, we include one dataset from each
179   setup, that we refer to as FLUXCOM-RS (from the RS setup) and FLUXCOM-MET (from the MET setup). To
180   choose the two datasets we analysed the pair-wise error correlations of all the products against in-situ
181   flux tower and selected the two that had the lowest pair-wise error correlation (and so were deemed least
182   dependent). In DOLCE V3, we include a dataset from the MET setup only.

183   Process-based Land Surface Evapotranspiration/Heat Fluxes algorithm (PLSH; Zhang et al., 2015):
184   Terrestrial ET is derived using an improved NDVI-based Penman-Monteith algorithm originally developed in
185   (Zhang et al., 2010). ET is regulated by a set of geophysical data from GIMMS and Vegetation Index and
186   Phenology along with radiative data from World Climate Research Programme/Global Energy and
187   Water-Cycle Experiment (WCRP/GEWEX) Surface Radiation Budget (SRB) and CERES along with other
188   meteorological observations data from the NCEP/DOE AMIP-II Reanalysis (NCEP2; Kanamitsu et al.,
189   2002).

190   Surface Energy Balance System (SEBS; Chen et al., 2019; Su, 2002): ET estimates are produced with the
191   revised Surface Energy Balance System (SEBS) algorithm in Chen et al. (2013; 2019). It uses
192   meteorological observations, ground heat flux, net radiation and canopy measurements collected from
193   flux tower sites, and NDVI and emissivity data from MODIS.

194   Surface Radiation Budget (SRB)-GEWEX (Vinukollu et al., 2011): ET is estimated based on the Penman-
195   Monteith equation. Input data sets include remote sensing data from AVHRR and MODIS,
196   meteorological data derived from the Variable Infiltration Capacity (VIC; Liang et al., 1994) land surface
197   model forced by PGF and radiative data from the NASA Global Energy and Water Exchanges (GEWEX)
198   Surface Radiation Budget Project (Stackhouse Jr et al., 2011).
199
200   It is clear that different parent datasets share forcing, parameterisations, and physical and empirical
201   assumptions. Therefore, they do not constitute entirely independent estimates. Furthermore, their error
202   correlation (when compared with data from 254 sites – details on these below), which can be used as a

203   measure of their dependence (Bishop and Abramowitz, 2013) is high (Fig. S2, correlation > 0.5),

204   reinforcing the potential for benefit using a weighting approach that can account for this redundancy.

205

206   Part of the high correlation is of course due to spatial heterogeneity and the scale mismatch between in-

207   situ and gridded data-sets —since individual site locations within a grid cell are likely biased with respect

208   to the (unknown) true grid cell averaged flux. While it might appear that a weighting approach that

209   accounts for error correlations between parent data-sets might be in danger of overfitting to error

210   correlation resulting from spatial heterogeneity, we have two mechanisms that ensure this is not a

211   concern for our final product. First, weights for each product are constructed over very large

212   spatiotemporal domains, i.e. more than 13000 space-time records as described below, so that the

213   (assumed stochastic) biases of individual sites relative to grid cell values are unlikely to influence weights

214   over a large sample. In fact, representativeness of point-scale measurement for the grid scale does exist

215   across all the flux tower sites as a whole, this has been verified byin (Hobeichi et al., (2018). Second, and

216   more categorical, all results here are presented out-of-sample, so that any overfitting will degrade,

217   rather than improve the results we present. More detail on this is presented below.

218

219   Given that most of the parent datasets provide ET information at a 0.25° or finer spatial resolution

220   (Table 1), it is possible to enhance the resolution of DOLCE from 0.5° to 0.25°. All the parent datasets are

221   resampled from their original spatial resolution to a common 0.25° grid using the nearest-neighbour

222   resampling method, and aggregated to monthly temporal scale before implementing the weighting

223   technique.

224

225

226   ### 2.1.2 Flux tower data

227   We use flux tower observations from a range of networks including Ameriflux (ameriflux.lbl.gov), tThe

228   Atmospheric Radiation Measurement (ARM; arm.gov), AsiaFlux (asiaflux.net), European Fluxes Database

229   (europe-fluxdata.eu), Fluxnet 2015, LaThuile Free Fair Use (fluxnet.fluxdata.org), Oak Ridge data

230   repository (daac.ornl.gov), OzFlux (ozflux.org.au), and data acquired through communication with

231   individual site principal investigators (PI). Particular efforts were made to establish connections with PIs

232   in regions where ET observations are scarce, including all areas outside North America, Europe and

233   Australia, particularly the MENA regions, Siberia, Central Africa and the Amazon basin. Our efforts and

234   communications with many PIs unfortunately failed to incorporate flux data from some of these regions

235   (excepting those that are already available from the cited networks). Before the quality control process

236   detailed below, we had obtained data from 366 flux tower sites.

237

238   The raw data consists of a composite of half hourly, daily and monthly records. We compute daily

239   averages from half-hourly records for days where at least 80% of half-hourly LE records are available.

240   Subsequently, we compute monthly averages from daily records for months where at least 80% of daily

241   LE records are available. In DOLCE V1 we applied a less strict quality control on the observational data in

242   which up to 50% of gap filling was allowed. The reason was that DOLCE V1 incorporated much fewer

243   observational data – sourced from Fluxnet 2015 and LaThuile Free Fair use only. In order to retain

244   enough observational data to constrain the weighting, it was necessary to make a trade-off between the

245   quality and the quantity of the data.

246

247   We also apply energy balance corrections to the monthly LE at all sites where monthly averages of the

248   other variables of the surface energy budget - net radiation (R_n), ground heat flux (G), and sensible

249   heat flux (H) - are available with the same high quality (quality flag > 80%). Corrections are carried out

250   independently for every monthly record. Where any of the other components of the energy budgets are

251   absent, latent heat measurements are used without any corrections~~directly~~. The energy balance

252   correction is applied as a Bowen Ratio (BR) based correction that distributes the energy budget residuals

253   among *H* and LE in such a way that their ratio is conserved. This is done under pre-defined constraints

254   that disallow large changes to be applied to LE. As a result of this, we accept the BR correction and use

255   the corrected LE (LE_{cor}) ~~LE_{cor} values~~ if the original monthly LE and LE_{cor} ~~the corrected LE (LE_{cor})~~

256   satisfy:

257
$$\begin{cases} \dfrac{LE_{cor}}{LE} \in \left[\dfrac{1}{2} - 2\right], & where \ LE \leq 30 \ W \ m^{-2} \\ LE_{cor} - LE \leq 20 \ W \ m^{-2}, & where \ LE \geq 30 \ W \ m^{-2} \end{cases}$$

258   ~~we accept the BR correction and use the corrected LE_{cor} values.~~ In DOLCE V1, we did not set a threshold

259   for LE adjustments, which resulted in LE being changed drastically in a few sites to offset errors in the

260   other energy balance components. If the BR correction does not meet the above criterion, we reject the

261   correction and try using a residual correction, which simply calculates LE as the residual term in the

262   energy balance equation, i.e. $LE_{cor} = R_n - H - G$. Similarly, we reject the residual correction if the

263   relation between LE and LE_{cor} above is not satisfied. In this case, we use the original monthly LE values

264   without correction. A simplified flowchart of these steps is displayed in Fig. S3 in the supplementary

265   material. A study by Paca et al. (2019) examined the changes to flux tower LE by three means of

266   correction~~s~~, and found that these on average differ by around 20 Wm-2 from one another. On this basis,

267   we expect that typically, the correction of flux tower LE should not exceed 20 Wm-2, unless errors in

268   other components of the budgets are propagating in the corrected ET. The rule for correcting small

269   fluxes and the condition in which each rule is applied (i.e. LE= 30 Wm-2) are in part subjective and in

270   another part based on a case by case assessment of changes induced to ET by the correction techniques,

271   and achieve a reasonable trade-off between data quality and availability.

272   In a further pre-processing step, if a site is located in close proximity to other sites such that they all sit

273   on the same 0.25° grid-cell, we use observational data from the site that is more representative of the

274   underlying grid-cell. Selecting the most representative site among these sites involves 1) identifying the

275   biome cover at each site; 2) computing the fraction of the grid area covered by each biome; the most

276   representative site is the one whose biome is more abundant in the underlying grid-cell (i.e. scores the

277   highest fraction of the total area). If all sites are equally representative of the underlying grid-cell, we

278   consider them as one site and we combine monthly LE from the sites by taking the average. We use the

279   high resolution 300 m - land cover maps from the European Space Agency (ESA; http://www.esa.int/)

280   downloaded from https://cds.climate.copernicus.eu/ to determine the biome types of neighbouring

281   sites and the corresponding grid-cells. This step has ensured that we are not matching a grid-cell with

282   inappropriate observational data. ~~This filtering reduced the number of employed sites in this study from~~

283 ~~366 sites to 260 sites (Fig. S1).~~ All the excluded sites are in Europe and North America. This filtering
284 along with the quality control measures described earlier reduced the number of employed sites in this
285 study from 366 sites to 260 sites (Fig. S1). Furthermore, we exclude 6 sites from the weighting, located
286 on flooded land area, wetlands or intensively irrigated land. As a result of this, the constraining
287 observational dataset used to derive DOLCE V2 includes 254 sites with a total of 13641 monthly records.
288
2.2. Methods

2.2.1 Weighting approach

291 The weighting technique is the same as that used in DOLCE V1 and was originally presented by Bishop and
292 Abramowitz (2013) and implemented for merging observational estimates by Hobeichi et al. (2018, 2019,
293 2020a). It consists of building a linear combination, μ, of the parent datasets that minimise
294 $\sum_{j=1}^{J}(\mu^j - y^j)^2$, where $j \in [1,J]$ are the monthly time-site records, y^j is the observed ET at the jth time-
295 site record. The linear combination $\mu^j = \sum_{k=1}^{K} w_k x_k^j$ is subject to the constraint that $\sum_{k=1}^{K} w_k = 1$,
296 where $k \in [1,K]$ represents the parent datasets and x_k^j is the value of the kth bias-corrected parent
297 dataset (i.e. after subtracting its mean bias relative to the all-site observational dataset) corresponding to
298 the jth time-site record. The analytical solution to this problem accounts for both the performance
299 differences between the parent datasets and their error covariance (Fig. S2), a proxy for dependence.
300 Further details on the merging technique can be found in Abramowitz and Bishop (2015) and Bishop and
301 Abramowitz (2013). The weighting approach is used to combine the global parent datasets separately on
302 different spatiotemporal subsets of the entire period and globe, using a tiered approach detailed in
303 section 2.2.3.
304
2.2.2 Computing uncertainty in ET

306 The ensemble dependence transformation process developed by Bishop and Abramowitz (2013) is used
307 to calculate the spatiotemporal uncertainty of DOLCE V2 and DOLCE V3. The process transforms the
308 global parent datasets to a new ensemble so that the variance of the transformed ensemble about the
309 derived hybrid ET estimate, μ, is constrained to be equal to the error variance of μ with respect to the
310 flux tower data, averaged over time and space (i.e. across all J records). We use the spread $\sqrt{\sigma^2}$ of the
311 transformed ensemble as the spatially and temporally varying estimate of uncertainty standard
312 deviation, which we will refer to as uncertainty. We refer the reader to Bishop and Abramowitz (2013)
313 for the derivation of this approach, and Hobeichi et al (2018) for its implementation in this context. The
314 spread $\sqrt{\sigma^2}$ of the transformed ensemble accurately reflects the uncertainty of μ in those grid-cells
315 where flux tower observations are available. This process ensures that the computed uncertainty
316 provides a better uncertainty estimate of the hybrid ET than simply using the spread of the parent
317 datasets.
318 One additional advantage of defining uncertainty in this way is that it should give an accurate upper
319 bound estimate of the likely discrepancy between the product and unseen ET measurements at a range
320 of spatial scales. That is, since it is based on the discrepancy of the final hybrid product and point-based

321 flux tower estimates, which are essentially at the extremes of spatial discrepancy, the discrepancy
322 between DOLCE and actual ET at any spatial scale greater than that of a tower footprint and smaller
323 than that of DOLCE should be less than this uncertainty estimate (noting however that this is the
324 estimated standard deviation of uncertainty, rather than a hard upper limit). In Section 2.2.56 below, we
325 detail the out-of-sample testing of this uncertainty estimate at the point scale.
326

2.2.3 Tiering of data set subsets in time and space to maximise coverage

328 To derive DOLCE V1 over the global land, we applied spatial tiering (using different subsets of parent
329 products in different regions to maximise spatial coverage). We now expand this approach to include
330 temporal tiering to improve the temporal reach of DOLCE. Collectively, the incorporated parent datasets
331 have a temporal cover over 1980 – 2018, but only a short common overlap during 2003-2007 in DOLCE
332 V2, and during 2003 – 2016 in DOLCE V3, and their spatial intersection does not cover the global land.
333 Therefore, to achieve a global land coverage from 1980 through 2018 without excluding any of their
334 parent products, it was necessary to build DOLCE V2 and DOLCE V3 from different subsets of parent
335 datasets in time periods and land regions depending on the availability of the parent datasets as shown
336 in Table 1. To this end, we consider 14 and 4 distinct temporal tiers in DOLCE V2 and DOLCE V3
337 respectively. For example, in DOLCE V2, tier 9 covers 2008 - 2012 and incorporates all datasets except
338 SRB-GEWEX. Tier 1 incorporates the least parent datasets, for the year 1980 (i.e. FLUXCOM-MET and
339 GLEAM3.3A), while tier 8 uses all the parent datasets and covers 2003 – 2007. Furthermore, within each
340 temporal tier, we consider three spatial sub-tiers, with each spatial sub-tier covering a part of the land.
341 These consist of (a) all land except Antarctica, Greenland and North Africa, (b) only Antarctica and
342 Greenland, (c) only North Africa. A similar spatial tiering approach was also applied in DOLCE V1. Other
343 spatial tiers, each consisting of a small number of grid cells were also considered where necessary to
344 ensure that no grid cell in DOLCE V2 or DOLCE V3 is missing ET data if a single parent is missing ET data
345 for that grid cell. As a result of the tiering approach, weighting is computed separately using a different
346 subset of parent data sets and site data in each tier, resulting in distinct spatiotemporal subsets of the
347 entire period. Note that in the results we below, we briefly examine the extent to which tiering results in
348 temporal discontinuities. Collectively, the hybrid estimates developed throughout the temporal tiers and
349 their spatial sub-tiers form DOLCE V2 and DOLCE V3 over the global land throughout 1980 – 2018. The
350 reduced number of temporal tiers in DOLCE V3 is to ensure that no temporal discontinuities occur
351 throughout the covered period, which otherwise would have reduced the suitability of DOLCE V3 for
352 trends analysis. In comparison, the incorporation of a larger ensemble of parent products in DOLCE V2 is
353 to derive an optimal ET product that minimises discrepancy with in-situ observations.

354
355

2.2.4 Weighting groups

357 Previous studies have found that the performance of a global product can vary with different climatic
358 circumstances, suggesting that separating the weighting into separate regions or other groupings might
359 well improve the results of the weighting overall (Ershadi et al., 2014; Hobeichi et al., 2018; Michel et al.,
360 2016). Grouped weighting simply involves dividing the time and/or space covered by a particular tier

361   into different subsets or groups (e.ge.g., with different climatic conditions), and then applying the

362   weighting technique separately for each group (within a single tier). We expect that grouped weighting

363   has the potential to improve weighting by accounting for the variation in performance of the parent

364   datasets over different climate or land conditions and can hopefully improve biases detected in DOLCE

365   V1. Hobeichi et al. (2018) tried to group flux tower sites based on their land cover type and computed

366   weights for each land cover type. However, this approach did not improve the results, whether grouping

367   by climate zone or aridity index, with the main reason being attributed to the small number of sites in

368   many groups. Despite the availability of 100 additional sites to constrain the weighting here compared

369   to Hobeichi et al., (2018), the ratio of the observational data to the number of parents has not improved

370   across several climate or land cover types for ~~this work~~DOLCE V2. We therefore investigate new

371   approaches to grouped weighting that allow sufficiently low group numbers to keep a reasonable

372   sample size in each of them, ~~,~~ including:

- Grouping by latitudinal zone: this is a simplification of grouping by climate type in which climates are aggregated into three latitudinal zones: (i) high latitudes (±60° poleward), (ii) mid-latitudes (±60° towards the subtropics ±40°), and (iii) tropics and sub-tropics (between -40° and 40°). In each zone we apply a separate weighting using the corresponding group of sites.
- Grouping by continents: Sites are naturally separated by continental boundaries and we might suspect that a particular ET product performs differently across continents. For instance, precipitation is involved in the derivation of many of the parent datasets, and has been found to have different fidelity over different continents (Hobeichi et al., 2020b).
- Grouping by hemisphere: Pan et al. (2020) found that ET estimates agree more in the Northern hemisphere than in the Southern hemisphere. Therefore, performing separate weighting in each hemisphere could be better than weighting across all global land.
- Grouping by seasons: Several studies have shown that the skill of ET datasets vary by seasons ~~(Jiménez et al. 2018; Long, Longuevergne, and Scanlon 2014; Mueller et al. 2011)~~ (Jiménez et al., 2018; Long et al., 2014; Mueller et al., 2011). To capture these differences, we implement grouping by seasons, and grouping by month (detailed below). 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~~December–May~~, and (2) monthly observations from sites located in the Southern hemispheres during the period December~~June~~–November~~May~~. The remaining observational data is used to constrain the weighting during the winter-spring combined season.
- Grouping by months: This is similar to grouping by seasons, the only difference is that the two groups are June–November and December–May, without accounting for the different seasonal phase between hemispheres.
- Grouping by ET regime and months: Land was classified into three distinct broad ET regimes (Fig. S4) according to two aspects of ET, mean annual total ET and within-year relative variability throughout 1980 – 2018, derived from GLEAM V3.5a, and using K-means unsupervised classification (MacQueen, 1967). We explain the classification method further in section 3.5.2. Different sets of weights were computed at each ET regime during June–November and December–May. Implementing weighting this way ensured that we account for performance

differences across different physical aspects of the land and seasons. Despite that observational data was divided into six distinct groups, the observational data available in each group was still appropriate to merge the four parent datasets of DOLCE V3. However, we found this grouped weighting strategy not appropriate for merging 11 parent datasets of DOLCE V2.

As an alternative to the grouping strategies, we also investigate if deriving a spatially varying bias correction within each tier could further improve the weighting. We describe the examined bias correction approaches and their effectiveness in the Supplementary Material.

2.2.5 Bias correction strategies

In DOLCE V1, we showed that part of the success of the weighting approach is due to the bias correction applied before the weighting. Within each tier, the bias correction is applied simply by adding the mean difference between a product and tower data uniformly to all values of a product before the weights are derived – it is constant in space and time for a given product within one tier. The grouping strategies detailed above examine the effect of considering different bias correction and weighting subgroups within each spatiotemporal tier, with groups divided by region (continents or latitudes) or/and seasons. As an alternative to the grouping strategies, we also investigate if deriving a spatially varying bias correction within each tier could further improve the weighting. A spatially varying bias correction might better capture the performance deficiencies of each the parent datasets.

To derive a global bias correction for a particular parent dataset within each tier, we first compute the mean bias at each flux tower site across all the time records within the tier. We then assign those ET bias values to the grid cells containing the sites. Finally, using the bias values at these grid cells, we extrapolate the bias field to the entire global land domain within the tier using several different extrapolation strategies, including inverse distance weighting (IDW), local polynomial interpolation and nearest neighbourhood. As with the different weighting groups, we test the effectiveness of each approach using out-of-sample tests, which we now describe.

2.2.5 Out-of-sample testing approach

To test the effectiveness of different weighting groups or bias-correction approaches, and assess which strategy offers the best performance, we use out-of-sample tests. To do this, we first divide the flux tower sites between the in-sample and out-of-sample groups by randomly selecting 25% of the sites as out of sample. The remaining sites form the in-sample training set are used to compute bias correction terms and weights for the parent datasets in each tier using the weighting technique without weighting groups (as adopted in DOLCE V1), and with each of the groups and bias correction strategies detailed above.in section 2.2.5 and S4 (supplementary material). In each case, these bias correction terms and weights are then applied to the parent datasets and compared to the out-of-sample sites to test efficacy of the clustering or bias correction approach employed. The process is repeated for each grouping or bias correction strategy to derive several hybrid ET datasets for each out of sample group of sites.

445 For each strategy, the test was repeated 1000 times with a different random selection of sites being out
446 of sample. The performance of each hybrid ET estimate was evaluated across five statistical metrics.
447 These were root mean squared error (RMSE), absolute standard deviation difference $|\sigma_{dataset} -$
448 $\sigma_{observation}|$, correlation, mean absolute deviation (i.e. mean(|dataset – observation|)) and median
449 absolute deviation (i.e. median(|dataset – observation|)). DOLCE V1 has not been included in this test
450 because its coarser spatial resolution (i.e. 0.5°) excludes many coastal sites and so significantly reduces
451 the observational data we could use in this analysis. The out-of-sample test is carried out over the
452 common period of availability of all the parent datasets i.e. 2003 – 2007 and 2003 – 2016 to enable
453 comparison of the out-of-sample performance of each approach with all of the 11 and 4 parent datasets
454 of DOLCE V2 and DOLCE V3 respectively.
455
456 We perform another out-of-sample experiment to test if the uncertainty estimate derived by the
457 successful grouping/bias correction strategy performs well out of sample. In this test, we first select a
458 site S, but instead of constraining the weighting using observed ET from this site, we compute the
459 weights and bias correction terms of the parent datasets by using all the sites except S (i.e. just one site
460 is out -of -sample). We then calculate the MSE of the derived hybrid ET against observations from all the
461 sites except S. We denote this value by $\text{uncertainty}_{\text{in-sample}}$, since it represents the uncertainty
462 estimate computed using the same observational dataset that we used to train the weighting. We also
463 calculate the MSE of the hybrid ET against the out-of-sample observations from S, and we denote this as
464 $\text{uncertainty}_{\text{out-sample}}$, since we perform the comparison against ET observations that have not been
465 used to train the weighting. We repeat this test for all the sites, and each time we calculate the ratio
466 $\frac{\text{uncertainty}_{\text{in-sample}}}{\text{uncertainty}_{\text{out-sample}}}$. In an ideal case, this ratio should equal to unity.
467
468
469 ## 3. Results and Discussion
470
471 ### 3.1 Out-of-Sample ~~Selection of a grouping strategy~~ Performance of DOLCE V2 and
472 DOLCE V3
473
473 We derive DOLCE V2.1 (Hobeichi, 2020) from 11 parent datasets by applying a grouped weighting by
474 months. As detailed in the Supplementary material (section S5), this approach achieves slightly better
475 out-of-sample performance than the other grouped weighting approaches in estimating ET (Fig. S6) and
476 in deriving more robust uncertainty estimates (Fig. S7). We recall that in grouped weighting by months,
477 the observational and gridded ET data are split into two groups, one covering the period June –
478 November and the other covering December – May. Weighting and bias correction is then implemented
479 in each group separately for each tier to create the subsets from which the hybrid ET product is derived.
480
480 ~~Figure 1 shows the out-of-sample performance of different grouping strategies (including no grouping)~~
481 ~~against parent datasets (left column). The performance results across all 1000 different random site~~
482 ~~samples are shown in a boxplot for each clustering method (yellow), non-clustered weighting (as per~~

483 ~~DOLCE V1, in magenta, labelled NO.GROUPING), and each parent dataset (purple). The hybrid ET~~
484 ~~estimates derived from grouping weighting are labelled LAT.ZONES, CONTINENTS, SEASONS, MONTHS,~~
485 ~~and HEMISPHERE, following the grouping approaches outlined above. The plots in the left column show~~
486 ~~that overall, the hybrid ET estimates outperform their parent datasets across all the performance~~
487 ~~metrics and in all clustering settings. The hybrid ET estimates derived by implementing spatially varying~~
488 ~~bias correction strategies failed to outperform the parent datasets in the out-of-sample site tests, and~~
489 ~~have been excluded from the plot (Fig. S3). To highlight the differences between the grouping strategies,~~
490 ~~we magnify the leftmost section of these plots in the right column of Fig. 1, and also show in red the~~
491 ~~median value of each boxplot. Results only change slightly across the grouping approaches, with the~~
492 ~~best results achieved by grouping weights by months. Despite the relatively small improvement offered~~
493 ~~by this strategy at the out-of-sample sites over the other grouping strategies, we derive DOLCE V2~~
494 ~~(Hobeichi, 2020) by applying a grouped weighting by months. Recall that in this approach, the~~
495 ~~observational and gridded ET data are split into two groups, one covering the period June – November~~
496 ~~and the other covering December – May. Weighting and bias correction is then implemented in each~~
497 ~~group separately for each tier to create the subsets from which the hybrid ET product is derived.~~

498 We derive DOLCE V3 (Hobeichi, 2021) from 4 parent datasets by applying a grouped weighting by ET
499 regimes and months. ~~The box plots in Fig. 2 show the ratio $\frac{Uncertainty_{in-sample}}{Uncertainty_{out-sample}}$ obtained across the~~
500 ~~different grouping techniques. Each boxplot represents this ratio from all sites out of sample and shows~~
501 ~~that over half of the data, the ratio ranges between 0.83 and 1.51, with a median very close to 1. This~~
502 ~~confirms that, overall, when the uncertainty estimates are computed out of sample, they are very~~
503 ~~similar to what they would have been if they were computed in sample. Also, the fact the shift in ratio is~~
504 ~~mostly towards values bigger than 1 rather than smaller than 1 indicates that $Uncertainty_{in-sample}$ is~~
505 ~~greater than $Uncertainty_{out-sample}$ so that uncertainty is overestimated rather than underestimated.~~
506 ~~Interestingly, the lower (0.86) and upper (1.46) quartiles achieved by grouping weighting by months are~~
507 ~~the closest to 1 than the other grouping techniques. This suggests that overall, grouping weighting by~~
508 ~~months is able to derive slightly more robust uncertainty estimates than the other techniques.~~

509 Both DOLCE V3 and DOLCE V2.1 outperform their parent datasets in the out-of-sample tests across all
510 performance metrics (Fig. S6 and Fig. S8). DOLCE V2.1 performs better than DOLCE V3 across all
511 performance metrics except Standard deviation difference as illustrated in Fig. S8. The overall better
512 performance of DOLCE V2.1 is expected given th~~at~~e more ET estimates contributing to the weighting. On
513 the other hand, DOLCE V2.1 has proven worse performance than DOLCE V3 in capturing variation in ET
514 observations since variability in ET should have decreased when the variations in individual products are
515 not temporally coincident.
516

517 ## 3.2 Comparison of DOLCE V2 and DOLCE V3 with their~~its~~ parent datasets

518 Figure 1~~3~~ displays the latitudinal means of each of ~~DOLCE V1,~~ DOLCE V2 and DOLCE V3 and their~~its~~
519 parent datasets computed over a common spatial ~~land~~ mask and common periods of ~~for~~ 2003 – 2007
520 and 2003 – 2016 in the case of DOLCE V2 and DOLCE V3 respectively. ~~We exclude Antarctica from the~~

521 ~~analysis due to the lack of reliable reference information on ET for validation.~~ The grey ribbon ~~in Fig. 3~~
522 represents the uncertainty of DOLCE V2 and DOLCE V3 in Fig. 1a and Fig. 1b respectively, defined by the
523 ± uncertainty standard deviation interval. ~~The pink shaded areas represent latitudinal domains where~~
524 ~~some parent datasets do not estimate ET for some parts of the land. In the white area where all the~~
525 ~~datasets have complete terrestrial coverage, t~~The uncertainty standard deviation of the two DOLCE ~~V2~~
526 products mostly contain~~s~~ the latitudinal variations of ~~its~~ their parent datasets with the exception of
527 FLUXCOM-RS which exhibits larger ET over the tropics and subtropics of the southern hemisphere
528 relative to DOLCE V2 (Fig. 1b). This containment should not be surprising since uncertainty estimates
529 should be robust for point-scale estimates. ~~Similarly, the middle pink area shows that DOLCE V2 agrees~~
530 ~~with its parent datasets except BACI, FLUXCOM-RS, FLUXCOM-MET and MOD16. These four datasets do~~
531 ~~not estimate ET over arid and semi-arid regions in north Africa, the middle east and central Asia, so it is~~
532 ~~expected that they exhibit larger ET averages over these latitudes.~~ Figure 1a shows that DOLCE V1
533 exhibits a slightly lower ET than DOLCE V2 in the tropics and sub-tropics. DOLCE V2 appears in the lower
534 end of the range of the other datasets from 60° poleward. All the datasets exhibit considerable
535 disparities over the mid-latitude south of -50°, where the contribution of the terrestrial ET comes mostly
536 from the lower Andes. The difference between DOLCE V2 and DOLCE V3 is smallest over the mid
537 latitudes of the Northern hemisphere where most of the flux tower sites are located, and is largest over
538 the tropics where very few observations are available. Also, both the number and the spread of parent
539 datasets is larger in DOLCE V2 which explains its larger uncertainty compared to DOLCE V3. The parent
540 datasets of DOLCE V3 are in general in the upper range of ET across all the different participating
541 products, which also explains why DOLCE V3 exhibits larger ET than DOLCE V2 throughout the land and
542 mostly over the tropics.
543
544 Figure ~~4~~ 2 shows the spatial distribution of differences in the ET mean between DOLCE V2 and each of its
545 parent datasets. We apply different spatiotemporal masks for each comparison based on parent dataset
546 coverage (Table 1). We also compute the climatological difference of DOLCE V2 with its predecessor
547 DOLCE V1 over 2000 – 2009. A similar plot showing the spatial distribution of differences in the ET mean
548 between DOLCE V3 and each of its parent datasets is provided in Fig. S9. Fig. S9 shows that DOLCE V3
549 exhibits higher ET than DOLCE V2.1 and DOLCE V1 over most of the land, particularly over the tropics
550 and the high latitudes. On the other hand, the climatological difference between DOLCE V3 and its
551 parent datasets show different spatial patterns, and the least climatological difference is between
552 DOLCE V3 and GLEAM V3.5B.
553
554 Over the temperate regions of the northern hemisphere, DOLCE V2 exhibits lower mean ET than all its
555 parents except SEBS. We have computed the mean bias of all these datasets relative to the
556 observational data available from sites located in these temperate latitudes. DOLCE V2 has a negligible
557 bias of 0.2 $W\ m^{-2}$ relative to the observational data. This bias results from a positive bias of 0.4 $W\ m^{-2}$
558 during June – November and a negative bias of -0.2 $W\ m^{-2}$ during December–May. All the parent
559 datasets except SEBS exhibit a positive bias that ranges between 2.7 and 11.4 $W\ m^{-2}$ and SEBS has a
560 negative mean bias of -3.4 $W\ m^{-2}$, that varies between -0.2 Wm^{-2} during December – May and -6.3
561 Wm^{-2} during June – November. We note that the bias relative to the in-situ observational datasets is
562 only indicative of the performance of the gridded datasets at the sites and do not necessarily represent

563   the actual mean bias over these regions. The discrepancy between DOLCE V2 and DOLCE V1 is relatively
564   small across all land.

565

566   Large differences between DOLCE V2 and FLUXCOM-RS are seen over the Congo and the Amazon basins,
567   southern Africa, and the Brazilian highlands. The mean climatological bias of FLUXCOM-RS relative to
568   observational data from these regions is 30 $W\ m^{-2}$. This large bias likely results from the lack of
569   sufficient data available to train the machine learning algorithm over climatically distinct biomes, which
570   made ET prediction less constrained. This bias did not appear in FLUXCOM-MET possibly because ET
571   prediction is based on a larger set of ~~-~~predictor variables~~The bias is likely to be the reason for the~~
572   ~~exceptionally large FLUXCOM-RS ET seen over the tropics in Fig. 3. On the other hand,~~ DOLCE V2 exhibits
573   a ~~much smaller~~relatively small ~~-~~bias ~~than FLUXCOM-RS that ~~rang~~es~~ between 2.6 $W\ m^{-2}$ during June–
574   November and 6.4 $W\ m^{-2}$ during December–May. In comparison, DOLCE V3 exhibits no significant~~-~~ bias
575   during June–November and a bias of 12.2 $W\ m^{-2}$ during December–May which is similar to the bias in
576   GLEAM V3.5B over these latitudes and seasons, and is less than the bias in the remaining parent
577   datasets (i.e. GLEAM V3.5A, FLUXCOM-MET, and ERA5
578   In general, there are apparent disparities in the patterns of climatological differences in the tropics
579   across all the maps. This results from the fact that global ET datasets exhibit large differences over the
580   tropics which has been highlighted previously (Paca et al., 2019; Pan et al., 2020), particularly over the
581   Amazon basin.

582

583   ~~For reference, we provide in global maps of the seasonal climatology of DOLCE V2 computed throughout~~
584   ~~1980 – 2018 in Fig. S5.~~

585

586   ## 3.3 Comparison of basin and continental ET with existing literature

587   We now compare DOLCE V2 and DOLCE V3 with annual mean ET aggregates over a range of river basins
588   documented in a recent study (Table 4 of Zhang et al., 2018). ET in this study - which we'll refer to as
589   CDR-ET- is derived by merging 10 available ET datasets into a hybrid ET which then receives corrections,
590   so that the surface water budget - established by derived hybrid estimates of the other hydrological
591   variables - is closed. Table 2 displays the mean annual ET aggregates in $mm\ year^{-1}$ across 20 river
592   basins calculated for DOLCE V2, DOLCE V3~~2~~ and CDR-ET over the common period 1984 – 2010. Our
593   results show that there is an overall agreement between DOLCE V2 and CDR-ET across all the non-
594   Siberian rivers where the difference in ET estimates is mostly around 10%. The agreement worsens over
595   the Arctic basins Indigirka, Kolyma, Lena, Northern Dvina, Yenisei and particularly over Olenik and
596   Pechora where the differences in ET estimates exceed 20%. Previous studies have reported large
597   uncertainties in the water fluxes over the Siberian basins (Lorenz et al., 2015) most likely due to the
598   absence of a proper representation of snow and permafrost dynamics (Candogan Yossef et al., 2012).
599   Interestingly, over the north American arctic basins Mackenzie and Yukon, DOLCE V2 and CDR-ET exhibit
600   much smaller relative differences than at their Siberian counterparts. DOLCE V3 exhibits higher ET than
601   DOLCE V2 and CDR-ET across the majority of the river basins, and particularly over the Arctic basins.
602   DOLCE V3 is within the range of its recently developed parent datasets which exhibit higher ET than the
603   old generation products such as SRB-GEWEX and SEBS incorporated in DOLCE V2.
604

605 We also compare DOLCE V2 and DOLCE V3 with continental annual means of ET shown by L'Ecuyer et al.
606 (2015). In their study, they derive a hybrid ET by merging three global datasets. Then, they adjust the
607 hybrid ET and its associated uncertainty by enforcing the physical constraints of the surface and
608 atmospheric water and energy budgets using a data assimilation technique (DAT). Our results show that
609 DOLCE V2 has smaller ET with larger associated uncertainties compared to those derived in L'Ecuyer et
610 al. (2015) (Table 3). The range of their ET estimate overlaps with the ~~upper~~ range of DOLCE V2 and
611 DOLCE V3 throughout all continents. In L'Ecuyer et al. (2015), the uncertainty estimates are originally
612 taken from the literature and are deemed constant across time and space, then these are reduced by
613 the DAT. ~~The uncertainty estimate of DOLCE V2 is firmly grounded in the spread of its parent ET datasets~~
614 ~~but is more robust than this spread alone, since this spread has been recalibrated so that the~~
615 ~~uncertainty of DOLCE V2 relative to the observational data is precisely the spread of the recalibrated~~
616 ~~parent datasets.~~ The uncertainty estimate of DOLCE, however, is firmly grounded in the discrepancy
617 between the gridded DOLCE product and in-situ tower data. The variance of this discrepancy is used to
618 recalibrate the variance of the parent datasets, which are then used to estimate uncertainty, allowing
619 spatiotemporally varying uncertainty estimate that is both consistent with the discrepancy between
620 DOLCE and surface observations while at the same time being spatially and temporally complete. This
621 process is detailed by Hobeichi et al (2018).

622
623 Finally, we compare DOLCE V2 with the ET component of Conserving Land Atmosphere Synthesis Suite
624 (CLASS; Hobeichi, 2019; Hobeichi et al., 2020a) which we denote as CLASS-ET. CLASS dataset comprises
625 coherent estimates of the surface water and energy budgets at the gridded monthly scale. CLASS-ET has
626 been derived by adjusting DOLCE V1 by enforcing the simultaneous closure of the surface water and
627 energy budgets using the same DAT as in L'Ecuyer et al. (2015), and can be therefore considered an
628 improved version of DOLCE V1. Table S3 displays the continental area weighted averages of DOLCE V2,
629 DOLCE V1 and CLASS-ET and the mean differences DOLCE V2 – DOLCE V1 and DOLCE V2 – CLASS
630 computed over a common time period 2003-2009, and ~~a~~ using a common spatial mask. We find that, in
631 general, DOLCE-V2 is closer to CLASS-ET (i.e. the improved version of DOLCE V1), than DOLCE V1.

632
633 ~~The average global land ET of DOLCE V2 during 1980 – 2018   is 37 $W\ m^{-2}$. This falls in the lower range~~
634 ~~of global ET climatology of 35 – 54 Wm^{-2} computed across 20 ET datasets during 1982 – 2011   in Pan~~
635 ~~et al. (2020)~~

636
637 ## 3.4 Performance of DOLCE V2 at flux sites

638 We now compare DOLCE V2 with ET measured at the 260 sites used ~~to derive it~~ in this study (Table S1).
639 We display two performance metrics - correlation and standard deviation - on a Taylor Diagram (Fig. ~~5~~3).
640 ~~RMSE has been excluded from the plot since the mean ET exhibited by DOLCE ET at a particular site does~~
641 ~~not necessarily equal the mean observed ET at that site.~~ All data has been normalised before computing
642 the statistical metrics so that the observational data at each site has a mean of zero and a standard
643 deviation of 1. Each coloured point summarises the performance statistics of DOLCE V2 at a single site.
644 The observational data is represented by a single "reference" point~~,~~ i.e., the hollow point at one on the
645 horizontal axis. The plot in Fig. ~~5~~3 shows that most of the coloured points lie close to the reference
646 point, indicating that DOLCE V2 is highly correlated with most of the observational data. Overall, Fig. ~~5~~3
647 shows good agreement with the observational datasets. Poor performance is seen over a small number

648   of sites. These are represented by points located outside the Taylor diagram area. Most of these sites
649   have less than one year of monthly records with several gaps, perhaps raising questions about
650   observational quality.
651   In a further analysis, we investigate whether the performance of DOLCE V2 is reduced over a particular
652   land cover type. For this purpose, we repeat Fig. 5̶3, but this time we colour-code the statistics points by
653   the land cover type of the sites they represent as shown in Fig. S̶6̶S10. The new plot does not reveal
654   clear links between the performance of DOLCE V2 and the biome types of the sites. Similarly, we could
655   not find performance links with the degree of representativeness of the site to the underlying grid-cell.
656   This is shown in Fig. S̶7̶ S11 where colours represent the degree of agreement between the land cover
657   type at the footprint of the tower site and the dominant land cover of the grid-cell containing the site.
658   As shown in Fig. S̶7̶S11, we carry out this analysis on the basis of three levels of agreement. These
659   include blue points representing sites whose land types match the dominant land types of the
660   underlying grid-cells; green points representing sites whose land types cover more than 25% of the
661   underlying grid-cells without being the dominant land cover at these grid-cells; and pink points
662   representing sites whose land types covers less than 25% of the underlying grid-cells.
663
664   Finally, Fig. S8 shows timeseries of DOLCE V2 and observed ET at a selection of sites from various climate
665   types. Site properties are shown in Table S1.
666

3.5 Changes in ET since 1980

3.5.1 Annual ET trends over the global land

669   We use DOLCE V3 to produce a long-term (1980 – 2018) map of trends in annual ET totals (Fig. 4̶6) as
670   proposed by Mann-Kendall (Kendall, 1948; Mann, 1945) using the Sen's slope method (Sen, 1968). We
671   use the uncertainty estimates associated with the ET fields and the confidence interval of the slope as
672   two confidence measures to filter out spurious trends. These confidence measures consider trends'
673   behaviour as reliable only if (i) the confidence interval of the slope does not encompasses a mix of
674   negative and positive values; and (ii) trends' slopes computed for multiple different random samples of
675   ET within the interval ET ± uncertainty s̶tandard d̶eviation agree in sign at least 90% of the time.

676   Unreliable trends occur in regions where ET uncertainty is relatively high, such as in c̶e̶n̶t̶r̶a̶l̶ north Africa
677   and Sahel, and in the high latitudes where ET observations are sparse or do not exist. Inconsistent trend
678   behaviour (CI includes positive and negative values) is found in regions that experienced long phases of
679   droughts and non-droughts during 1980-2018, mainly in Australia, or a succession of drought and wet
680   events, mainly in southern United States and m̶o̶s̶t̶ ̶o̶f̶ the Amazons basin (Marengo et al., 2018). As a
681   result of this, a general long trend in ET is not identified in these regions. Miralles et al. (2014) report
682   that these changes in ET over these regions reflect El-Niño-La-Niña cycle. Similarly, we have not
683   detected clear long trends in southern South America and eastern and southern Africa. This partially
684   agrees with the study of Pan et al. (2020) where their figure 8 shows no ET trend in eastern Africa, and
685   no agreement on the sign of trend between the participating datasets has been found in southern South
686   America. Figure 6̶ 4 indicates that ET has increased̶t̶e̶n̶s̶i̶f̶i̶e̶d̶ over most of the northern latitudes which
687   has been highlighted in many studies (e.g. Miralles et al. 2014; Pan et al. 2020; Zhang et al. 2016), and
688   declined in western United States, e̶a̶s̶t̶e̶r̶n̶ ̶I̶n̶d̶i̶a̶,̶ ̶m̶o̶s̶t̶ ̶o̶f̶ ̶M̶a̶d̶a̶g̶a̶s̶c̶a̶r̶,̶ ̶a̶n̶d̶ ̶p̶a̶r̶t̶s̶ ̶o̶f̶ ̶t̶h̶e̶ ̶E̶t̶h̶i̶o̶p̶i̶a̶n̶

689 ~~highland~~central Africa and South Amercia. Unfortunately, given the absence of adequate in-situ
690 observations that cover a long enough period to establish trends analysis, it is difficult to validate the
691 identified trends directly.

692 In further analysis, we ~~examine whether~~verify that the spatiotemporal tiering adopted in DOLCE ~~V2~~ V3
693 has not resulted in temporal discontinuities. Figure ~~7~~ 5 illustrates the annual average line plot of the
694 area weighted mean of continental ET exhibited by DOLCE ~~V2~~V3. The vertical dashed lines mark the
695 beginning of a new tier ~~(see Table 1~~, i.e. in 1981, 2003 and 2017~~)~~. While the line plot does shows some
696 marked changes, these do not coincide with changes in tiers, and rather ~~we do not believe these reveal~~
697 ~~a signal of temporal discontinuity, as most of the strong changes in ET that coincide with changes in tiers~~
698 ~~also~~ coincide with extreme events, and are specific to the continents where these events occurred. For
699 instance, in Australia, ET shows high mean annual total in three very wet years 2000, 2010 and 2011,
700 and low levels throughout 2001 – 2009 during the millennium drought. Additionally, ~~For instance, the~~
701 ~~drop in ET in South America in 2016 is explained by an unprecedented drought over tropical South~~
702 ~~America (Erfanian et al., 2017). Similarly, the drop in ET in Africa is caused by several droughts occurring~~
703 ~~in many African regions since 2016. In Australia,~~ the decline in ET since 2017 is caused by severe
704 droughts that developed across most of Australia.
705
706 ## 3.5.2 ET regimes

707 To understand changes in ET across wet and dry regions, we classify land into 6 distinct dry and wet ET
708 regimes according to two aspects of ET: annual averages and within-year relative variability derived
709 from DOLCE V3. We apply K-means clustering (MacQueen, 1967) - an unsupervised machine learning
710 algorithm known for its outstanding efficiency in clustering data – by implementing the K-Means
711 function and the least squares quantisation method (Lloyd, 1982) using R software. K-Means identifies K
712 centroids (i.e. imaginary values representing the centre of the clusters) and assigns each data point to
713 the cluster of the nearest centroid using – in this paper - the least squares quantisation method. For
714 each grid cell, we compute 1) the average of the annual total ET across 39 years (1980-2018); and 2)
715 within-year relative variability climatology by temporally averaging the relative standard deviation of
716 monthly ET calculated over a year and across all years. These have been used as input features for the
717 unsupervised classification. After trial and error, we find that the global land can be adequately
718 classified into six distinct regimes that include three dry and three wet regimes. According to centroids
719 values (Table S4), we label the six regimes from driest to wettest and we list the proportion of the land
720 covered by each regime~~list the associated ET climatology and variability respect~~ively: (i) very low ET
721 with high variability (~~3 mm,~~ 1~~64~~%), (ii) low ET with high variability (~~207mm,~~ 34~~10~~%) ~~-~~, (iii) mild low ET
722 with medium variability (~~371 mm,~~ 22~~7~~%), (iv) mild high ET with medium variability (~~631mm,~~ 13~~5~~%), (v)
723 high ET with low variability (~~913mm,~~ 8~~3~~%), and (vi) very high ET with low variability (~~1221,~~ 7~~1~~%). Figure 6
724 ~~8~~ displays the spatial distribution of the 6 ET regimes.

725 We compare the derived ET regimes map with the modified Köppen climate (KC) classification map by
726 Chen and Chen (2013). We find that each KC class overlaps with only one ET regime with only two
727 exceptions (Table 4): i) Land characterised by a 'Dry Steppe Hot arid' (coded BSh in KC) climate belongs
728 the 'Mild low ET with medium variablity regime', but in two regions, the Indian Deccan plateau and
729 Argentinean Gran Chaco low forests, where the climate is BSh, the ET regime is 'Mild high ET with

730 medium variability'; ii) Regions with a 'Mild temperate Fully humid Hot summer' climate (coded Cfa in
731 KC) overlaps with the 'Mild high ET with medium variability' regime in coastal regions, and to the 'Very
732 high ET with low variability' regime in inland regions. These two KC classes (i.e. BSh and Cfa) are shown
733 in bold in Table 4. Overall, ET-regimes defined in this paper provide an efficient way to aggregate the KC
734 classes in less varied classes. This is not surprising knowing that KC classes are developed based on the
735 empirical relationship between climate and vegetation, and that ET links the water, energy (climate) and
736 carbon (vegetation) budgets.
737
738 ### 3.5.3 Global annual trends across the ET regimes
739 We now explore annual trends in mean ET exhibited in each ET regime during 1980-2018. First, we
740 calculate the annual ET total climatology and ET relative variability climatology spatially averaged across
741 each regime separately, then we compute the trends in yearly ET as above (i.e. using Mann-Kendall and
742 the Sen's slope methods). Figure ~~9~~ 7 illustrates trends' results for the dry regimes (V.L.ET, H.variability,
743 L.ET, H.variability and M.L.ET, M.Variability) and the wet regimes (M.H.ET, M.variability, H.ET,
744 L.variability and V.H.ET, L.variability). Across all regimes except the wettest one, trends in yearly ET total
745 are upward as indicated by the positive signs of both the slopes and their complete confidence intervals.
746 The strongest trends occur in the ~~'M.L.ET, M.Variability' and the~~ 'M.H.ET, M.variability' regime~~s~~ at a rate
747 ~~s 1.8~~0.6 $mm\ year^{-1}$ ~~and 1.78 $mm\ year^{-1}$ respectively~~, while the slowest trend occurs in the 'V.L.ET
748 H.variability' regime where ET is in general low. In the wettest ET regime 'V.H. ET, L. variability', while
749 the slope of the trend is positive, its confidence interval contains mixed positive and negative values.
750 This suggest that the tendency for increasing ET in the wettest ET regime is not robust. Our results
751 indicate that decreasing ET trends observed in some regions oppose the consistent positive trends
752 across the majority of ET clusters.
753 We repeat the same analysis for all ~~the 5~~ the participating parent datasets that span at least 30 years.
754 Sen's slope of the trends and their confidence interval (computed at the 95% confidence level) are
755 presented in Table 5. As noted earlier, ~~T~~trends' behaviour is deemed inconclusive when the CI
756 encompasses negative and positive values~~, in which case trends are considered unreliable~~. These are
757 presented with regular (as opposed to bold) typeface and are exhibited by FLUXCOM-MET in all regimes
758 except the driest. ERA5-land shows downward trends in the 'M.H.ET, M.variability' and 'H.ET,
759 L.variability' regimes. Both GLEAM 3.~~3A~~5A and PLSH show upward ET trends in all regimes, with the
760 exception of GLEAM which shows no reliable trends in the driest and wettest ET regime~~s~~. Differences
761 exist in the magnitude of trends ~~as DOLCE V2 shows in general the strongest trends~~ across the majority
762 ~~of the~~the products and the regimes. As in DOCLE ~~V2~~V3, the strongest trends in GLEAM 3.~~3A~~5A occur in
763 the ~~'M.L.ET, M.Variability' and the~~ 'M.H.ET, M.variability' regime at a rate 0.5 $mm\ year^{-1}$~~s'~~. Finally, the
764 slopes of DOLCE V3 trends are within the range of slopes of trends in available ET products.
765
766 There are of course some notable limitations to the approach we have taken here, some of which were
767 previously discussed in Hobeichi et al. (2018). First, the weighting approach adopted here relies heavily
768 on flux tower observations, which can suffer from a range of technical issues (Burba and Anderson,
769 2010; Fratini et al., 2019), as well as temporal gaps during particular weather conditions such as
770 extremes (Van Der Horst et al., 2019), which can affect our results. Next, unresolved land surface
771 processes in the parent datasets due for example to the absence of a proper representation of snow and

772 permafrost dynamics, or the heterogeneity of the land surface are likely to lead to uncertain ET
773 estimation in DOLCE V2 and DOLCE V3, since it each of these isis only a combination of its parent data
774 sets. This applies particularly in regions where observations are scarce or do not exist.
775
776

4. Conclusions

778 This work presents two new hybrid ET datasets DOLCE V2.1 and DOLCE V3. e derivation of a new hybrid
779 ET dataset allowed us to examine historical trends in ET and their robustness to observational
780 uncertainty. The dataset, DOLCE V2,The new datasets are is publicly available and was the result of
781 several key improvements over theirits predecessor, incorporating more parent products in DOLCE V2.1,
782 more in-situ data, testing a range of alternative implementations of its weighting and bias correction
783 approach, increased spatial resolution, and coverings a longer time period. The incorporation of a large
784 ensemble of parent datasets in DOLCE V2.1 allowed us to derive a more optimal ET product that can be
785 used to benchmark global ET estimates. In comparison, the reduced number of parent datasets in
786 DOLCE V3 minimised temporal tiering and ensured that no temporal discontinuities occur throughout
787 the covered period. This allowed us to examine historical trends in ET and their robustness to
788 observational uncertainty. Despite the observationally constrained approach to defining uncertainty, we
789 found robust ET trends across most areas of the land surface, enough to present a clear signal in almost
790 eachmost of the ET climate regimes we examined. These trends indicate a global increase in land
791 derived ET between 1980 and 2018. This contrasts with other gridded ET products that did not
792 incorporate the same degree of observational constraint in either their mean field or uncertainty
793 estimates, and demonstrates the usefulness of this long-term hybrid ET dataset.
794

5. Data Availability

796
797 DOLCE V2.1 -dataset (Hobeichi, 2020) is publicly available in NetCDF-4 format and can be freely
798 downloaded from the NCI data catalogue at http://dx.doi.org/10.25914/5f1664837ef06.;
799
800 DOLCE V3 dataset (Hobeichi, 2021) is publicly available in NetCDF-4 format and can be freely
801 downloaded from the NCI data catalogue at https://doi.org/10.25914/606e9120c5ebe.
802

6. Competing interests.

804
805 The authors declare that they have no competing interests.
806

7. Acknowledgment

808
809 The authors acknowledge the support of the Australian Research Council Centre of Excellence for
810 Climate Extremes (CE170100023). This research was undertaken with the assistance of resources and

811 services from the National Computational Infrastructure (NCI), which is supported by the Australian
812 Government. We thank Franklin (Pete) Robertson (NASA Marshall Space Flight Center) for his valuable
813 contribution to DOLCE V3. This work used eddy covariance data acquired and shared by the FLUXNET
814 community, including these networks: AmeriFlux, AfriFlux, AsiaFlux, CarboAfrica, CarboEuropeIP,
815 CarboItaly, CarboMont, ChinaFlux, Fluxnet-Canada, GreenGrass, ICOS, KoFlux, LBA, NECC, OzFlux-TERN,
816 TCOS-Siberia, and USCCC. The FLUXNET eddy covariance data processing and harmonization was carried
817 out by the ICOS Ecosystem Thematic Center, AmeriFlux Management Project and Fluxdata project of
818 FLUXNET, with the support of CDIAC, and the OzFlux, ChinaFlux and AsiaFlux offices. Data were also
819 obtained from the Atmospheric Radiation Measurement (ARM) Program sponsored by the U.S.
820 Department of Energy, Office of Science, Office of Biological and Environmental Research, Climate and
821 Environmental Sciences Division. This works used data sourced from Terrestrial Ecosystem Research
822 Network (TERN) infrastructure, an Australian Government NCRIS enabled project; the Oak Ridge
823 National Laboratory Distributed Active Archive Center (ORNL DAAC); the Land Cover project of the ESA
824 Climate Change Initiative. We would like to thank all the principal investigators that authorised us to
825 download site data from the European Fluxes Database, and all the research institutes that made
826 publicly available and/or hosted the gridded ET datasets used in this study.
827
8. Tables
829
830 *Table 1: Spatial and temporal coverage and original resolution of the global ET datasets (at the time of analysis) used to develop*
831 *DOLCE V2.1 and DOLCE V3. DOLCE V2.1 was derived from 11 datasets and 14 temporal tiers. DOLCE V3 was derived from 4*
832 *datasets and 4 temporal tiers i.e. (1) 1980, (2) 1981 – 2002, (3) 2003 – 2016, (4) 2017 - 2018 The first column shows the number*
833 *of temporal tier.*

	Time period	BACI	ERA5-land	FLUXCOM-MET	FLUXCOM - RS	GLEAM 3.3A	GLEAM 3.3B	MOD16	PML	PLSH	SEBS	SRB-GEWEX
DOLCE Version		V2.1	V2.1 V3	V2.1 V3 (1980-2016)	V2.1	V2.1 V3 (GLEAMV3.5A & B)		V2.1	V2.1	V2.1	V2.1	V2.1
Tier Excluded Land domain		Antarctica Greenland North Africa		Antarctica Greenland North Africa	Antarctica Greenland North Africa			Antarctica Greenland North Africa	Antarctica Greenland			
	Original resolution	0.5° half hourly	0.1° hourly	$\frac{1°}{12}$ monthly	$\frac{1°}{12}$ monthly	0.25° monthly	0.25° monthly	0.05° monthly	0.5° monthly	$\frac{1°}{12}$ monthly	0.05° monthly	0.1° 3-hourly
1	1980			●		●						
2	1981		●	●		●			●			
3	1982 – 1983		●	●		●			●	●		
4	1984 – 1999		●	●		●			●	●		●
5	2000 1,2&3		●	●		●		●	●	●		●
6	2000 (4 – 12)		●	●		●		●	●	●	●	●
7	2001 – 2002	●	●	●	●	●		●	●	●	●	●

8	2003 – 2007	•	•	•	•	•	•	•	•	•	•	•
9	2008 – 2012	•	•	•	•	•	•	•	•	•	•	
10	2013	•	•	•	•	•	•	•			•	•	
11	2014	•	•	•	•	•	•	•	•			•	
12	2015		•	▲	•	•	•	•			•	
13	2016 – 2017 (1 – 6)		•				•	•			•	
14	2017 (7 – 12) – 2018		•				•	•				

834
835 *Table 2: Mean annual ET aggregates in $mm\ year^{-1}$ across 20 river basins calculated for DOLCE V2, DOLCE V3 and CDR-ET*
836 *(Table 4 of Zhang et al., 2018) over a common period 1984 – 2010. CDR-ET is derived by merging 10 available ET datasets into a*
837 *hybrid ET which then receives corrections, so that the surface water budget established by derived hybrid estimates of the other*
838 *hydrological variables is closed.*
839

Basin	CDR-ET 1984 – 2010	DOLCE V2 1984 – 2010	DOLCE V3 1984 – 2010
Amazon	1153	1167	1314
Amur	295	309	421
Columbia	331	340	436
Congo	1045	1084	1160
Danube	503	451	550
Indigirka	138	107	231
Indus	277	323	365
Kolyma	167	132	243
Lena	245	185	283
Mackenzie	241	214	333
Mississippi	577	513	555
Murray-Darling	411	419	445
Niger	401	456	427
Northern Dvina	324	232	376
Ob	323	245	357
Olenek	174	108	237
Paraná	892	854	856
Pechora	244	166	276
Yenisei	265	216	325
Yukon	175	158	261

840

Table 3: Annual continental averages of ET $(W\ m^{-2})$ and its standard deviation uncertainty calculated for DOLCE V2, DOLCE V3 and developed in (L'Ecuyer et al., 2015) over a common period 2000 – 2009~~10~~. In (L'Ecuyer et al., 2015), ET is derived by merging three global datasets, and then adjusted by enforcing the physical constraints of the surface and atmospheric water and energy budgets.

continent	ET± uncertainty ~~$(W\ m^{-2})$~~ (L'Ecuyer et al., 2015) ~~2000 – 2010~~	ET± uncertainty ~~$(W\ m^{-2})$~~ DOLCE V2 ~~2000 – 2010~~	ET± uncertainty DOLCE V3
Africa	45 ± 3	40 ± 17	39 ± 13
Australia	27 ± 3	28 ± 16	28 ± 13
Eurasia	33 ± 3	30 ± 13	34 ± 13
North America	33 ± 6	28 ± 12	32 ± 12
South America	77 ± 4	73 ± 23	76 ± 19

Table 4: correspondence between ET-regimes derived here and Köppen climate classes derived in (Chen and Chen, 2013. Text in bold fontface indicates that the Köppen climate is associated with more than one ET regime.

ET regimes	Köppen climate classes (Chen and Chen, 2013)
Very low ET with high variability	Polar (Tundra/Frost) Dry Desert (Hot/Cold) arid
Low ET with high variability	Snow Fully humid Cold summer/Cool summer Snow Dry summer Cool summer Snow Dry winter Cold summer Dry Steppe Cold arid Dry Desert Hot arid/Cold arid Mild temperate Dry summer Cool summer Mild temperate Dry summer Warm summer
Mild low ET with medium variability	Snow Fully humid (Hot/Warm summer) Snow Dry winter (Hot/Warm/Summer) **Dry Steppe Hot arid** Mild temperate Dry summer Hot summer Mild temperate Fully humid Warm summer
Mild high ET with medium variability	**Dry Steppe Hot arid (observed only in the Indian Deccan plateau and Argentinean Gran Chaco low forests)** **Mild temperate Fully humid Hot summer (observed in inland regions)** Mild temperate Dry winter (Hot/Warm summer) Tropical Dry summer
High ET with low variability	**Mild temperate Fully humid Hot summer**/Warm summer (**observed in coastal regions)** Tropical Dry winter
Very high ET with low variability	Tropical Fully humid Topical Monsoon

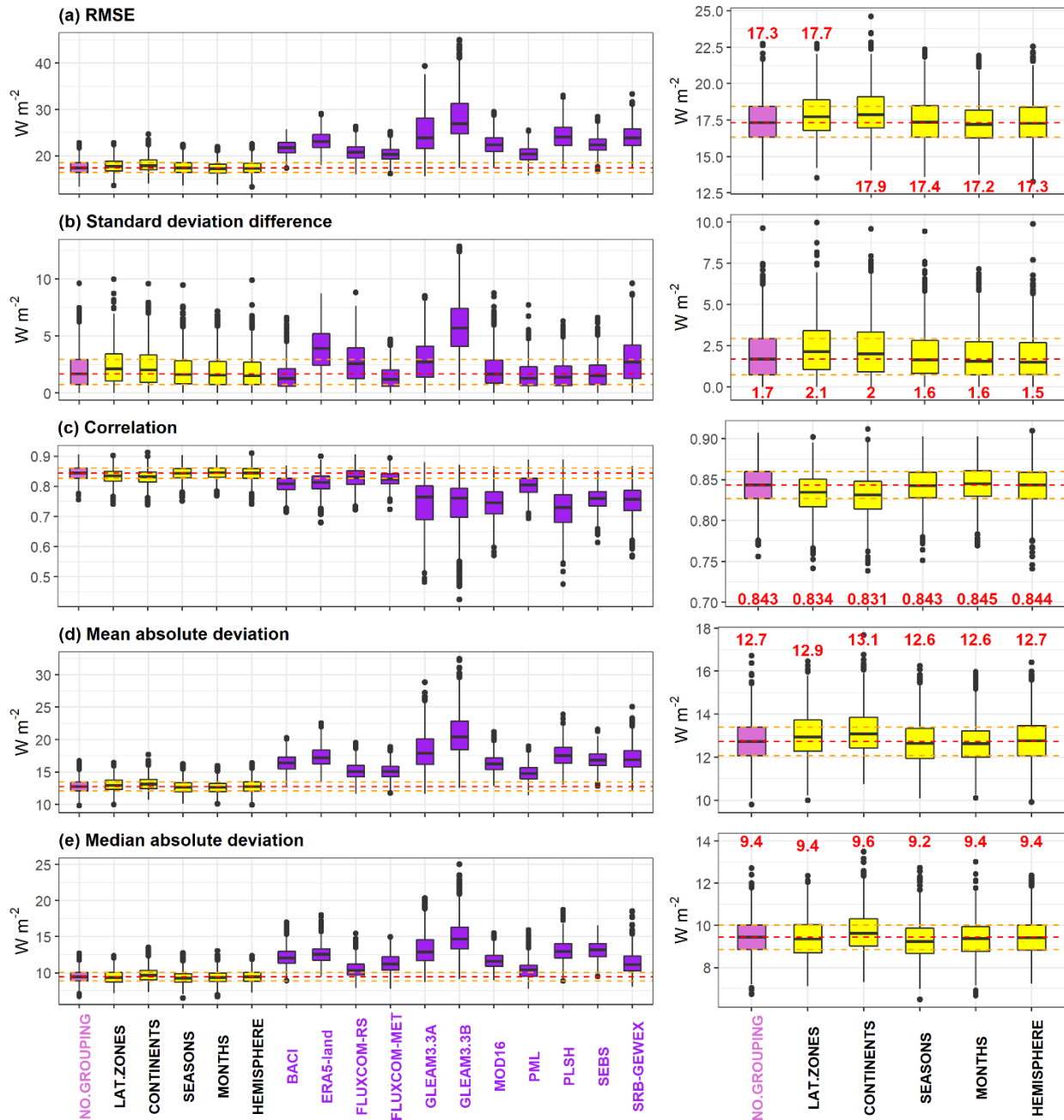
Table 5: Trends in yearly ET total (mm year^{-1}) spatially averaged across each ET regime calculated for DOLCE V3~~2~~ and ~~its~~ participating parent~~s~~ datasets that have time-span of more than 30 years. The text shows slopes of the trend line and their confidence interval calculated at the 95% confidence level, bold text indicates that the trend is reliable since the confidence interval is strictly positive or negative.

Dataset and time span	V.L.ET, H.variability	L.ET, H.variability	M.L.ET, M.Variability	M.H.ET, M.variability	H.ET, L.variability	V.H.ET, L.variability
DOLCE ~~V2~~V3 1980-2018	0.1 [0, 0.2]~~0.3 [0.1, 0.5]~~	0.2 [0, 0.4]~~0.9 [0.5, 1.2]~~	0.3 [0.1, 0.6]~~1.8 [1.4, 2.1]~~	0.6 [0.4, 0.8]~~1.7 [0.9, 2.5]~~	0.1 [0, 0.3]~~1.3 [0.4, 2.3]~~	0.2 [-0.1, 0.6]~~1.0 [0.3, 2.1]~~
ERA5-land 1981-2018	-0.1 [-0.2, 0.01]~~-0.1 [-0.2, 0.01]~~	-0.1[-0.4, 0.2]~~0.04 [-0.2, 0.3]~~	-0.05 [-0.3, 0.2]	-0.6 [-0.9, -0.3]~~-0.5 [-0.9, -0.2]~~	-0.5 [-0.8, -0.2]~~-0.8 [-1.1, -0.5]~~	0.04 [-0.2, 0.3]~~-0.08 [-0.3, 0.1]~~
FLUXCOM-MET 1980-~~2014~~2016	-0.01 [-0.03, 0]~~-0.01 [-0.07, 0.04]~~	0.04 [-0.06, 0.2]~~0.1 [-0.1, 0.3]~~	0.08 [-0.03, 0.2]~~0.1 [-0.05, 0.2]~~	-0.04 [-0.1, 0.1]~~0.05 [-0.1, 0.2]~~	0.03 [-0.1, 0.1]~~-0.04 [-0.2, 0.1]~~	0.1 [-0.1, 0.3]~~-0.1 [-0.3, 0.1]~~
GLEAM 3.~~5~~3A 1980-2018	0.01 [-0.1, 0.1]~~0.2 [-0.1, 0.4]~~	0.3 [0.02, 0.5]~~0.7 [0.5, 1.0]~~	0.5 [0.2, 0.7]~~1.2 [1.0, 1.5]~~	0.5 [0.2, 0.7]~~1.3 [1, 1.6]~~	0.3 [0.02, 0.5]~~0.6 [0.4, 0.9]~~	0.2 [-0.05, 0.6]~~0.3 [-0.1, 0.8]~~
PML 1981-2012	-0.1 [-0.3, 0.1]~~-0.1 [-0.3, 0.1]~~	0.4 [0.04, 0.7]~~0.4 [0.1, 0.7]~~	1 [0.5, 1.4]~~1.0 [0.5, 1.4]~~	0.2 [-0.2, 0.6]~~0.2 [-0.1, 0.6]~~	0.3 [-0.3, 0.8]~~0.08 [-0.4, 0.5]~~	-0.4 [-1.2, 0.5]~~-0.2 [-0.9, 0.6]~~
PLSH 1982-2013	0.16 [0.1, 0.2]~~0.1 [0.03, 0.1]~~	0.4 [0.2, 0.6]~~0.4 [0.2, 0.6]~~	1.2 [0.7, 1.6]~~1.0 [0.6, 1.5]~~	1.3 [0.8, 1.8]~~1.4 [0.9, 1.9]~~	1.5 [0.8, 2.1]~~1.5 [0.9, 2.1]~~	0.9 [0.5, 1.4]~~0.9 [0.5, 1.5]~~

854

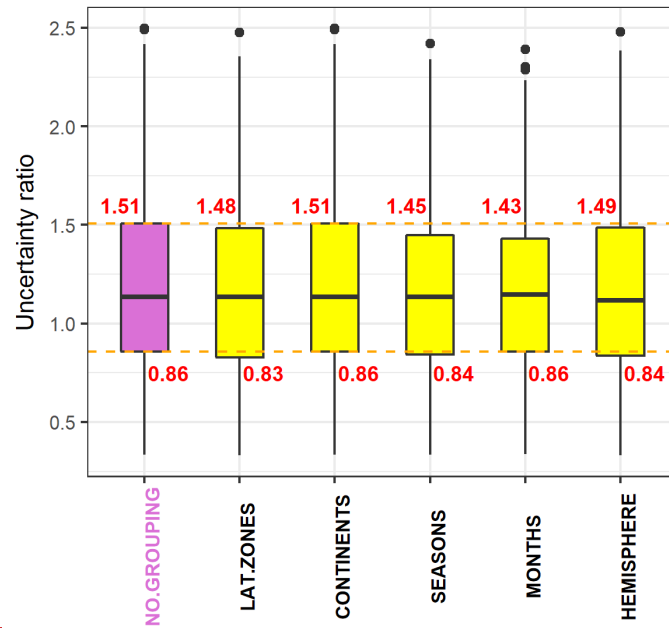
855

856

857

9. Figures

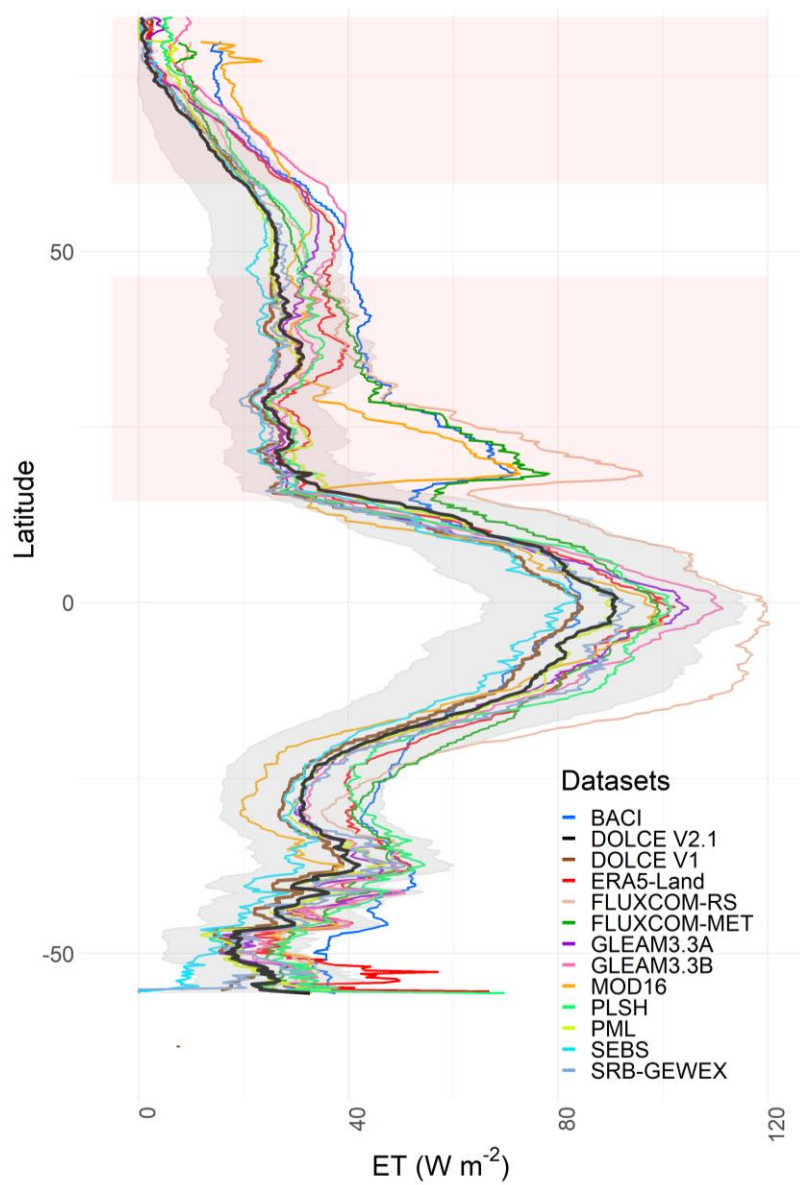
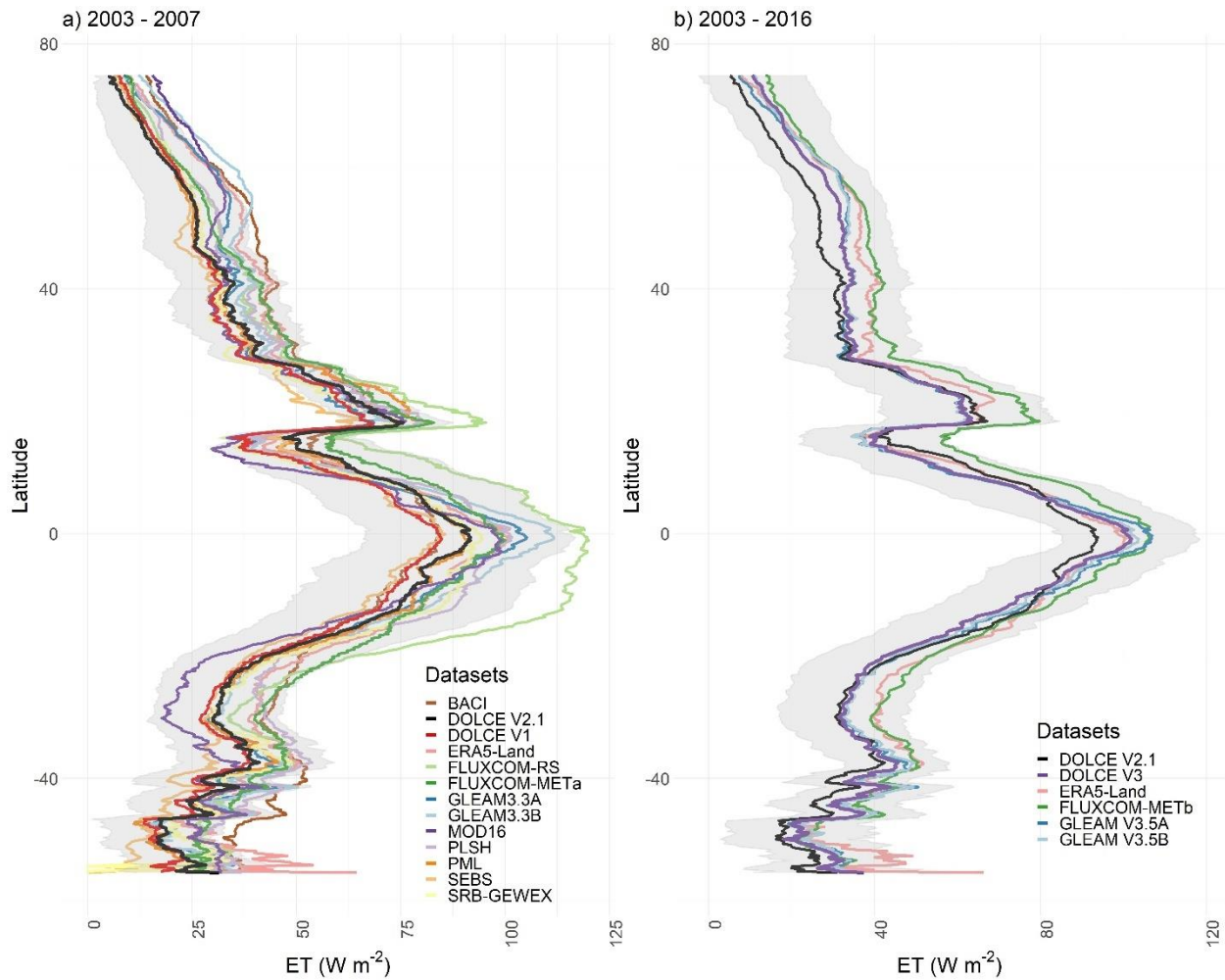
858
859 *Figure 1: Results of the out of sample test across five metrics of performance, (a) RMSE, (b) CORRELATION (c) SD difference , (d)*
860 *Mean Absolute Deviation , and (e) Median Absolute Deviation. Box plots represent spread over 1000 different selections of out-*
861 *of-sample sites. Different clustering methods (yellow) include: no clustering (NO.GROUPING; shown in magenta and horizontal*
862 *dashed lines), by latitude (LATZONES), by continents (CONTINENTS), by seasons (SEASONS), by months (MONTHS), and by*
863 hemisphere (HEMISPHERE), the red text marks the median values. Performance comparison with each of the parent datasets is
864 shown in purple.

865

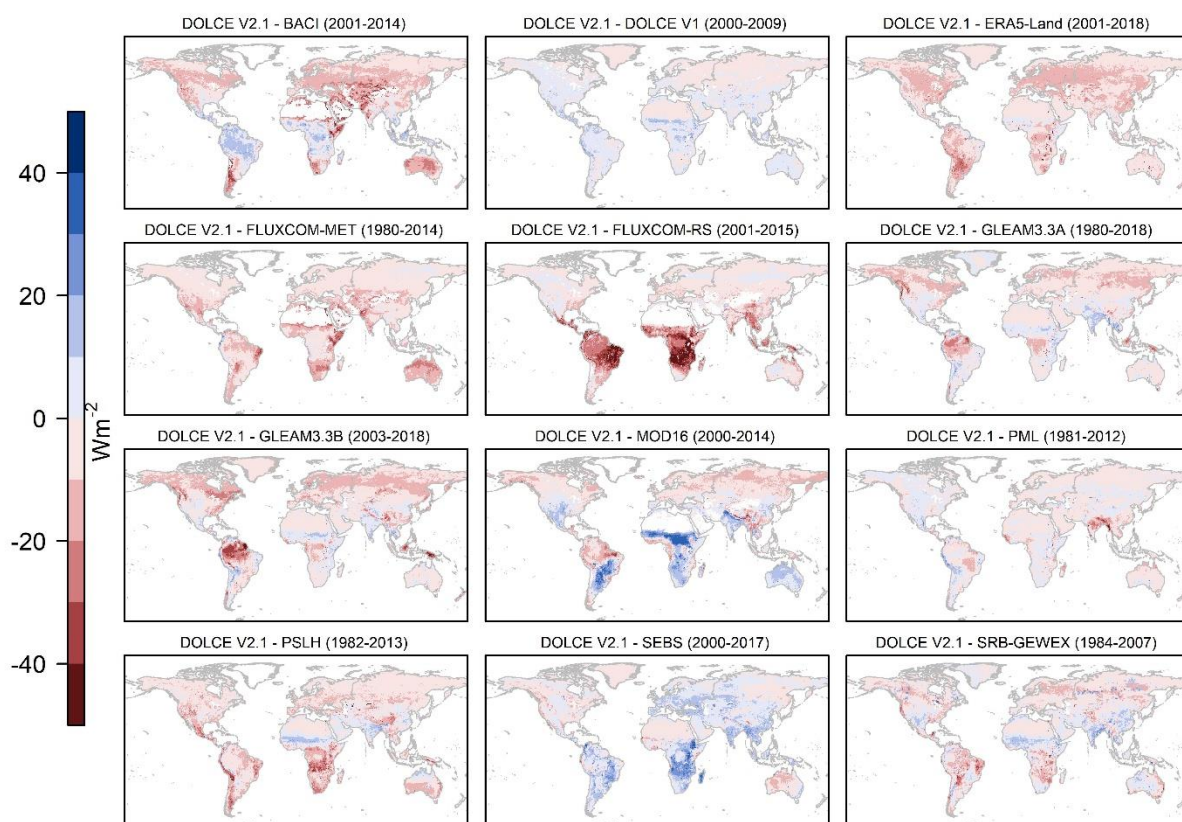
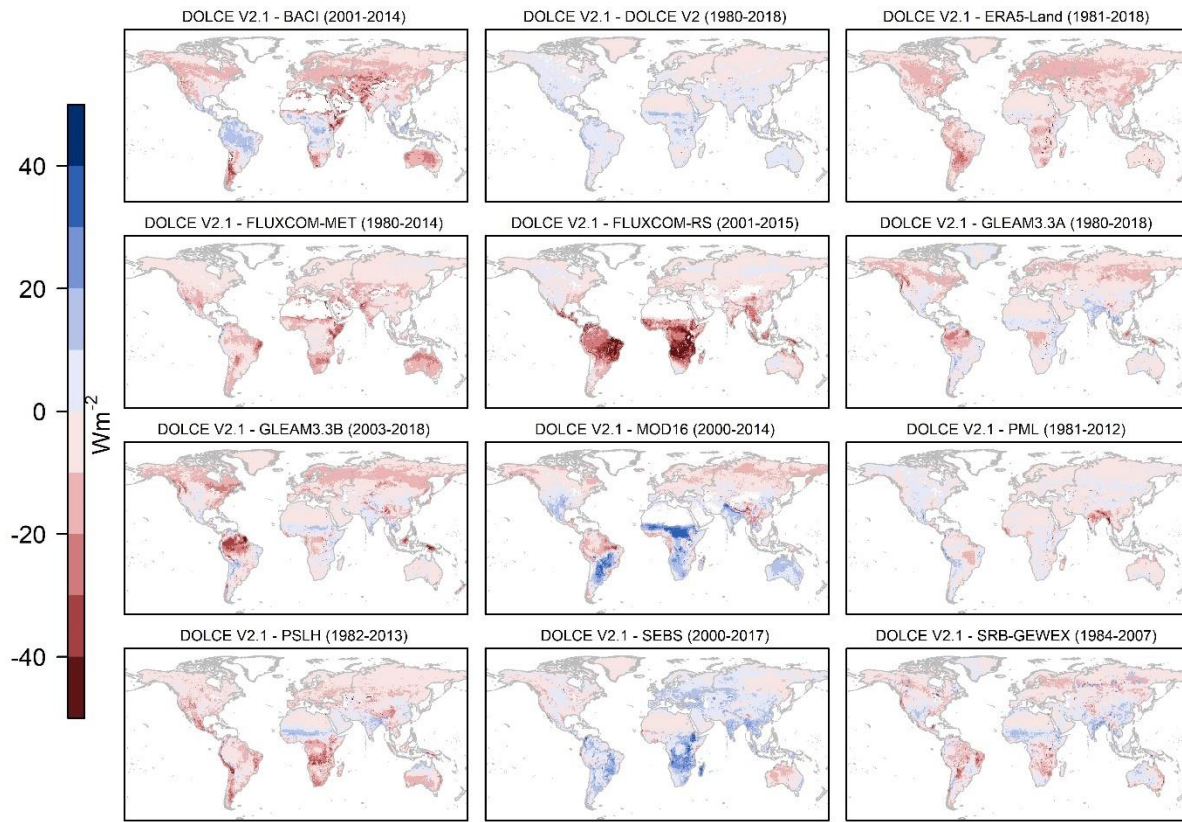
866

867 *Figure2: Box and whisker plots displaying the ratio* $\dfrac{Uncertainty_{in-sample}}{Uncertainty_{out-sample}}$*, computed for each site using the clustering methods*

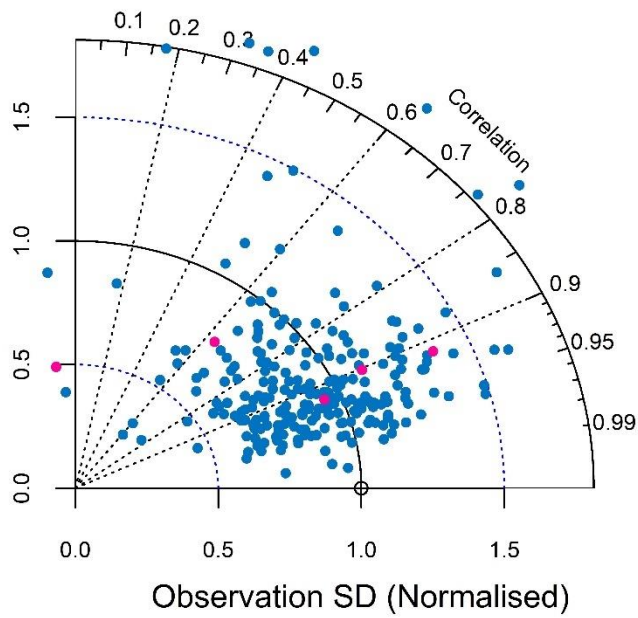
868 *defined in Sect. 2.2.4. Labeling and colors are as in Fig. 1. Red text marks the value of the upper quantile (75%) and lower*

869 *quantile (25%).*

871

872 *Figure 1̶3̶: (a) Latitudinal means of DOLCE V2 and its parent datasets computed over a common period 2003–2007, and a*
873 *common spatial mask. (b) Latitudinal means of DOLCE V3 and its parent datasets computed over a common period 2003–2016,*
874 *and common spatial mask. The grey ribbon represents the values of DOLCE ± uncertainty̶T̶h̶e̶ ̶g̶r̶e̶y̶ ̶r̶i̶b̶b̶o̶n̶ ̶r̶e̶p̶r̶e̶s̶e̶n̶t̶s̶ ̶t̶h̶e̶*
875 *u̶n̶c̶e̶r̶t̶a̶i̶n̶t̶y̶ ̶s̶t̶a̶n̶d̶a̶r̶d̶ ̶d̶e̶v̶i̶a̶t̶i̶o̶n̶ ̶o̶f̶ ̶D̶O̶L̶C̶E̶ ̶V̶2̶.̶ ̶T̶h̶e̶ ̶p̶i̶n̶k̶ ̶s̶h̶a̶d̶e̶d̶ ̶a̶r̶e̶a̶s̶ ̶r̶e̶p̶r̶e̶s̶e̶n̶t̶ ̶l̶a̶t̶i̶t̶u̶d̶i̶n̶a̶l̶ ̶d̶o̶m̶a̶i̶n̶s̶ ̶w̶h̶e̶r̶e̶ ̶s̶o̶m̶e̶ ̶p̶a̶r̶e̶n̶t̶ ̶d̶a̶t̶a̶s̶e̶t̶s̶*
876 *h̶a̶v̶e̶ ̶g̶a̶p̶s̶ ̶i̶n̶ ̶s̶o̶m̶e̶ ̶p̶a̶r̶t̶s̶ ̶o̶f̶ ̶t̶h̶e̶ ̶l̶a̶n̶d̶ ̶a̶s̶ ̶s̶h̶o̶w̶n̶ ̶i̶n̶ ̶T̶a̶b̶l̶e̶ ̶1̶. DOLCE V1 and DOLCE V2 are included in (a) and (b) respectively for*
877 *comparison. FLUXCOM-METa and FLUXCOM-METb are two different datasets from the FLUXCOM-MET setup.*

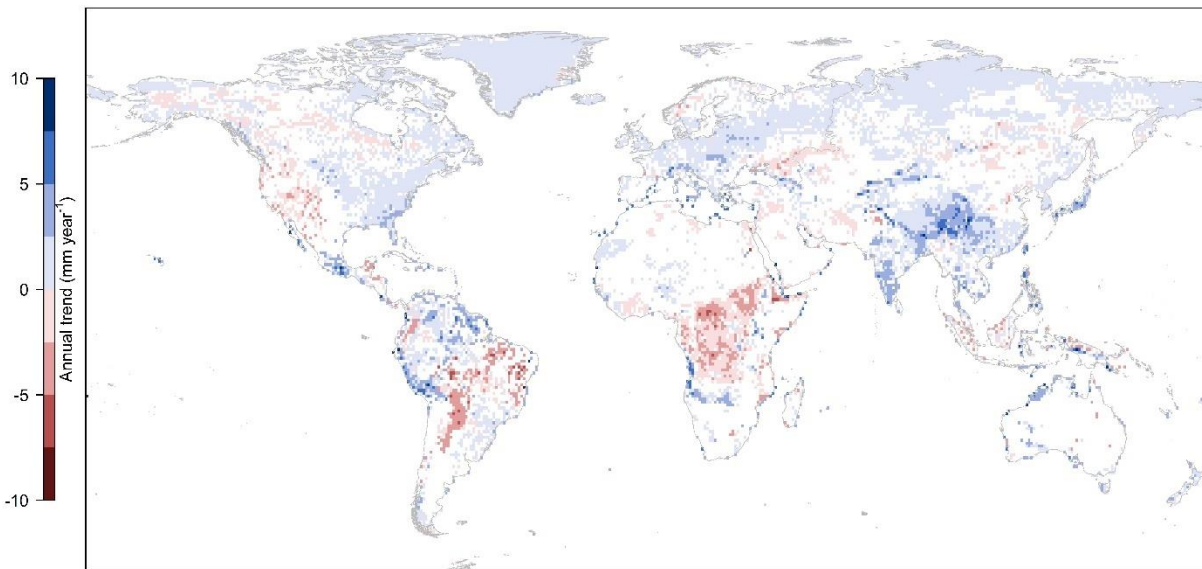
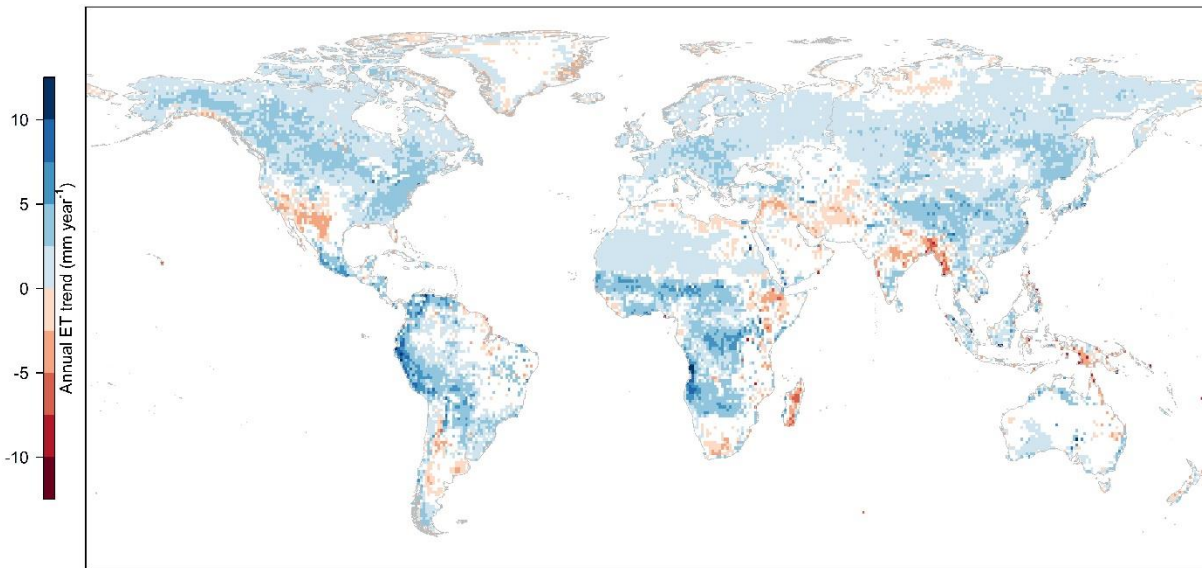
878

879

880
881 *Figure 24: Spatial distribution of differences in ET* ~~mean~~ *climatology between DOLCE V2 and each of its parent datasets and*
882 *DOLCE V1. Different spatiotemporal masks are applied for each comparison based on the spatiotemporal coverage of DOLCE V2*
883 *and the other datasets.*

884

885 *Figure 3̶5̶: Taylor Diagram displaying two performance metrics i.e. correlation and standard deviation of DOLCE V2 relative to*
886 *normalised observational data presented by a hollow point (reference point) at one unit on the x-axis. Pink points represent*
887 *performance statistics scored at sites located on wetlands, flooded plain or intensively irrigated areas.*

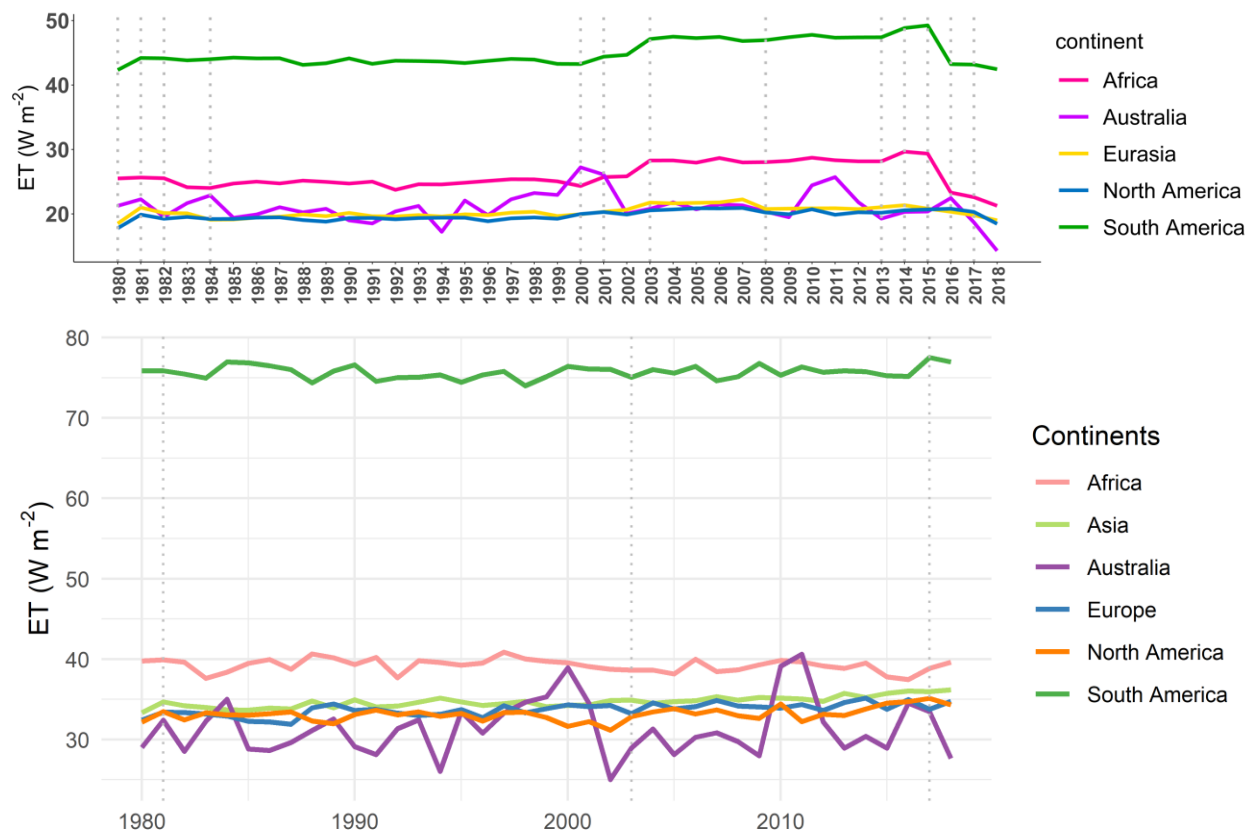
888

889

890

891 *Figure 6̶4̶: Spatial pattern of ET climate trends in DOLCE V3̶2̶ over 1980 – 2018 derived a̶s̶ using p̶r̶o̶p̶o̶s̶e̶d̶ ̶b̶y̶ Mann-Kendall and*
892 *Sen's slope methods. Grid cells in white correspond to unreliable ET trends because G̶r̶i̶d̶ ̶c̶e̶l̶l̶s̶ ̶i̶n̶ ̶w̶h̶i̶t̶e̶ ̶d̶o̶ ̶n̶o̶t̶ ̶e̶x̶h̶i̶b̶i̶t̶ ̶r̶e̶l̶i̶a̶b̶l̶e̶*
893 *t̶r̶e̶n̶d̶s̶ ̶a̶s̶ ̶i̶n̶d̶i̶c̶a̶t̶e̶d̶ ̶b̶y̶ ̶t̶h̶e̶ ̶i̶m̶p̶l̶e̶m̶e̶n̶t̶e̶d̶ ̶c̶o̶n̶s̶i̶s̶t̶e̶n̶c̶y̶ ̶m̶e̶a̶s̶u̶r̶e̶s̶.̶ (i) the confidence interval of the slope encompasses a mix of*
894 *negative and positive values; or (ii) trends' slopes computed for multiple different random samples of ET within the interval ET ±*
895 *uncertainty do not agree in sign.*

896

897

898

899 *Figure* ~~75~~: ~~Annual average line plot for area weighted mean of continental ET.~~ *Annual average line plot of the area weighted*
900 *mean of continental ET exhibited by DOLCE V3. The vertical dashed lines mark the beginning of a new tier in 1981, 2003 and*
901 *2017*

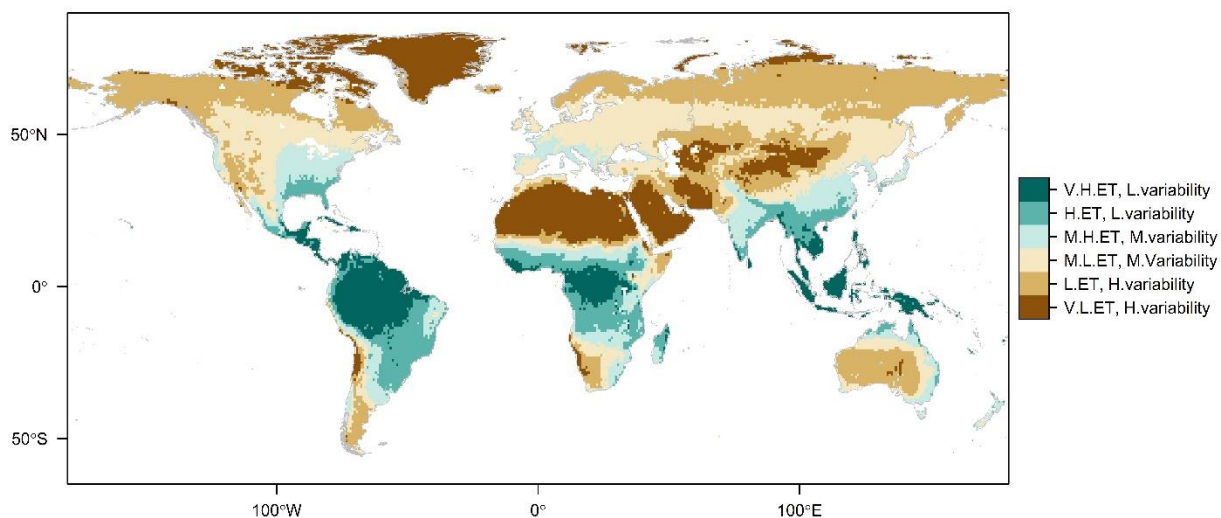
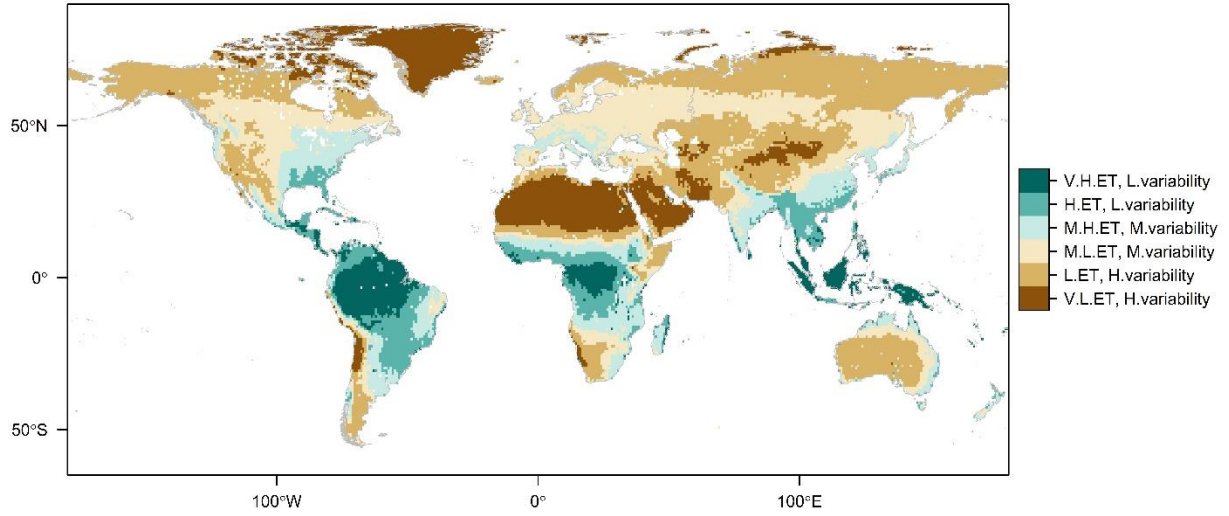
902

903

904

905

906

907

908 *Figure ~~86~~: Classification of the land into 6 distinct dry and wet ET regimes using K-means unsupervised classification based on*
909 *DOLCE V3 annual ET mean and within-year relative variability both computed for 1980-2018. The six ET regimes are labelled*
910 *from direst to wettest as very low ET with high variability (V.L.ET, H. variability), (ii) low ET with high variability (L.ET, H.*
911 *variability) , (iii) mild low ET with medium variability (M.L.ET, M. variability), (iv) mild high ET with medium variability (M.H.ET,*
912 *M. variability), (v) high ET with low variability (H.ET, L. variability), and (vi) very high ET with low variability (V.H.ET, L.*
913 *variability).~~Spatial distribution of ET regimes based on ET means and seasonal variations computed for 1980-2018.~~*

914

915

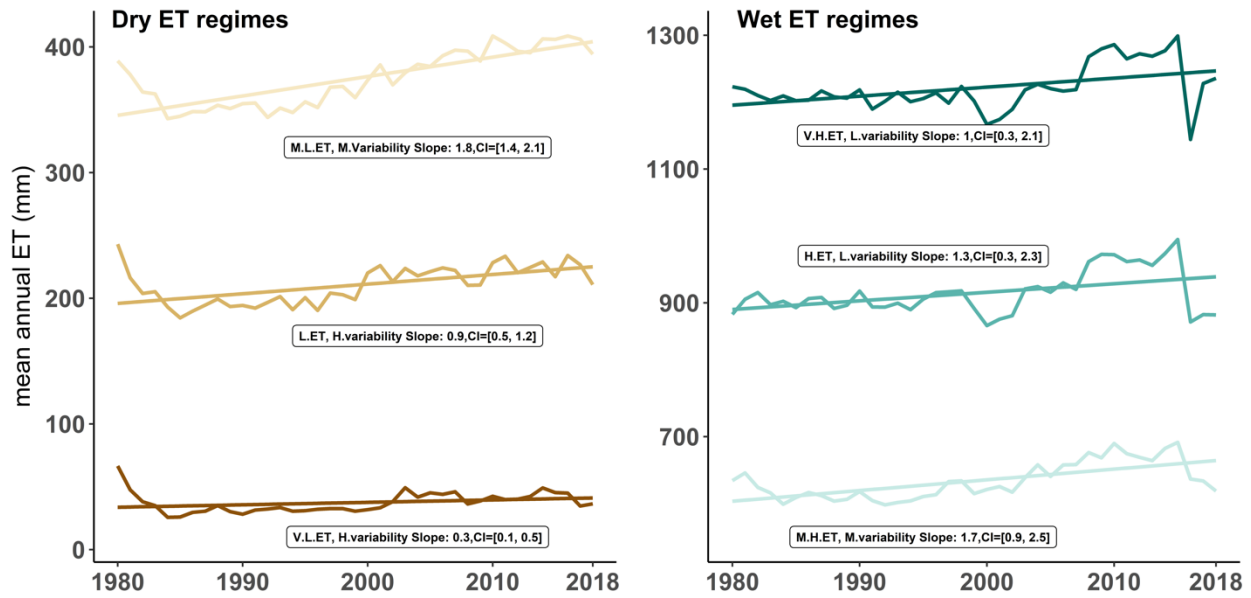
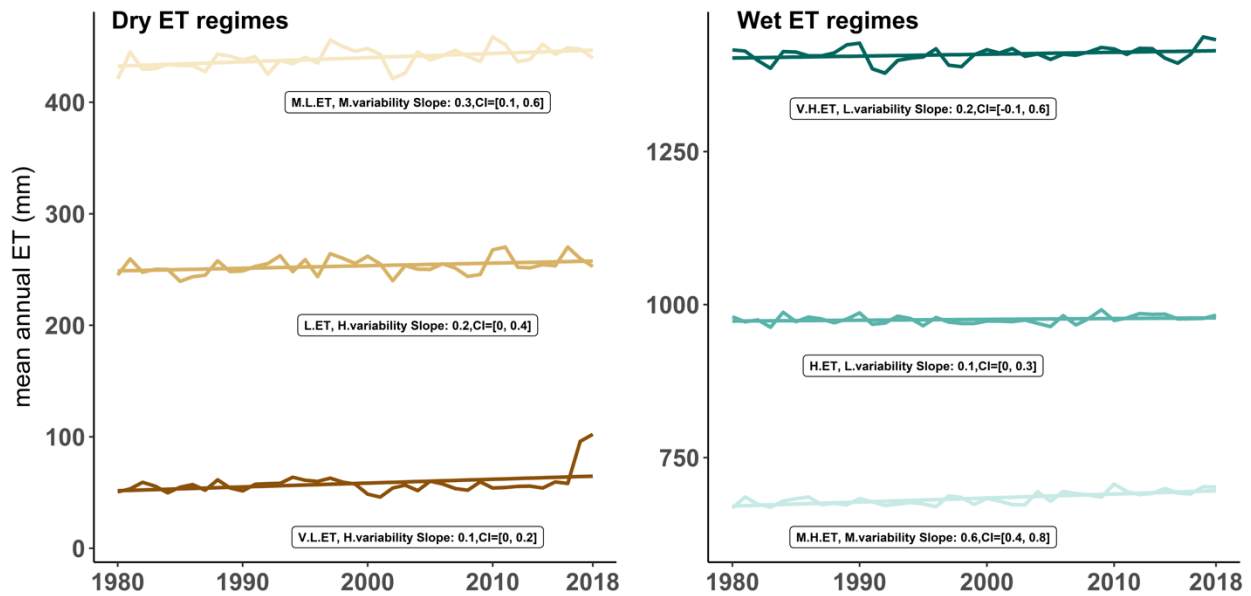



Figure ~~9~~7: Trends in mean ~~global~~ annual ET total computed for the dry and wet ET regimes during 1980-2018. Slopes and confidence intervals ~~(CI)~~ are computed using Mann-Kendall and the Sen's slope methods~~at the 95% significance level~~. The spatial distribution of the ET regimes is illustrated in Fig. 6.

10.   References

Abramowitz, G. and Bishop, C. H.: Climate model dependence and the ensemble dependence transformation of CMIP projections, J. Clim., 28(6), 2332–2348, doi:10.1175/JCLI-D-14-00364.1, 2015.

Abramowitz, G. ., Herger, N., Gutmann, Ethan; Hammerling, D. and Knutti, Reto; Leduc, Martin; Lorenz, Ruth; Pincus, Robert; Schmidt, G. A.: ESD Reviews: Model dependence in multi-model climate ensembles: weighting, sub-selection and out-of-sample testing, Earth Syst. Dynam, 10(1), 91–105, doi:10.5194/esd-10-91-2019, 2019.

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

Bishop, C. H. and Abramowitz, G.: Climate model dependence and the replicate Earth paradigm, Clim. Dyn., 41(3–4), 885–900, doi:10.1007/s00382-012-1610-y, 2013.

Bodesheim, P., Jung, M., Gans, F., Mahecha, M. D. and Reichstein, M.: Upscaled diurnal cycles of land-Atmosphere fluxes: A new global half-hourly data product, Earth Syst. Sci. Data, 10(3), 1327–1365, doi:10.5194/essd-10-1327-2018, 2018.

943   Burba, G. B. P. G. to E. C. F. M. P. and W. E. for S. and I. A. and Anderson, D.: A Brief Practical Guide to
944   Eddy Covariance Flux Measurements: Principles and Workflow Examples for Scientific and Industrial
945   Applications, LI-COR Biosciences., 2010.
946   Candogan Yossef, N., Van Beek, L. P. H., Kwadijk, J. C. J. and Bierkens, M. F. P.: Assessment of the
947   potential forecasting skill of a global hydrological model in reproducing the occurrence of monthly flow
948   extremes, Hydrol. Earth Syst. Sci., 16(11), 4233–4246, doi:10.5194/hess-16-4233-2012, 2012.
949   Chen, D. and Chen, H. W.: Using the Köppen classification to quantify climate variation and change: An
950   example for 1901–2010, Environ. Dev., 6, 69–79, 2013.
951   Chen, X., Su, Z., Ma, Y., Yang, K., Wen, J. and Zhang, Y.: An improvement of roughness height
952   parameterization of the Surface Energy Balance System (SEBS) over the Tibetan plateau, J. Appl.
953   Meteorol. Climatol., 52(3), 607–622, doi:10.1175/JAMC-D-12-056.1, 2013.
954   Chen, X., Massman, W. J. and Su, Z.: A column canopy-air turbulent diffusion method for different
955   canopy structures, J. Geophys. Res. Atmos., 124(2), 488–506, 2019.
956   Dawdy, D. R., Lichty, R. W. and Bergmann, J. M.: A rainfall-runoff simulation model for estimation of
957   flood peaks for small drainage basins, US Government Printing Office., 1972.
958   Erfanian, A., Wang, G. and Fomenko, L.: Unprecedented drought over tropical South America in 2016:
959   Significantly under-predicted by tropical SST, Sci. Rep., 7(1), doi:10.1038/s41598-017-05373-2, 2017.
960   Ershadi, A., McCabe, M. F., Evans, J. P., Chaney, N. W. and Wood, E. F.: Multi-site evaluation of
961   terrestrial evaporation models using FLUXNET data, Agric. For. Meteorol., 187, 46–61,
962   doi:10.1016/j.agrformet.2013.11.008, 2014.
963   Feng, F., Li, X., Yao, Y., Liang, S., Chen, J., Zhao, X., Jia, K., Pintér, K. and McCaughey, J. H.: An Empirical
964   Orthogonal Function-Based Algorithm for Estimating Terrestrial Latent Heat Flux from Eddy Covariance,
965   Meteorological and Satellite Observations, PLoS One, 11(7), e0160150,
966   doi:10.1371/journal.pone.0160150, 2016.
967   Fisher, R. A. and Koven, C. D.: Perspectives on the future of Land Surface Models and the challenges of
968   representing complex terrestrial systems, J. Adv. Model. Earth Syst., 12, 1–24,
969   doi:10.1029/2018ms001453, 2020.
970   Fratini, G., Sabbatini, S., Ediger, K., Riensche, B., Burba, G., Nicolini, G., Vitale, D. and Papale, D.:
971   Characterization of Eddy Covariance flux errors due to data synchronization issues during data
972   acquisition, in Geophysical Research Abstracts, vol. 21., 2019.
973   Hamed Alemohammad, S., Fang, B., Konings, A. G., Aires, F., Green, J. K., Kolassa, J., Miralles, D., Prigent,
974   C. and Gentine, P.: Water, Energy, and Carbon with Artificial Neural Networks (WECANN): A statistically
975   based estimate of global surface turbulent fluxes and gross primary productivity using solar-induced
976   fluorescence, Biogeosciences, 14(18), 4101–4124, doi:10.5194/bg-14-4101-2017, 2017.
977   Han, D., Wang, G., Liu, T., Xue, B.-L., Kuczera, G. and Xu, X.: Hydroclimatic response of
978   evapotranspiration partitioning to prolonged droughts in semiarid grassland, J. Hydrol., 563, 766–777,
979   2018.
980   Herger, N., Abramowitz, G., Knutti, R., Angélil, O., Lehmann, K. and Sanderson, B. M.: Selecting a climate
981   model subset to optimise key ensemble properties, Earth Syst. Dyn., 9(1), 135–151, doi:10.5194/esd-9-
982   135-2018, 2018.
983   Hobeichi, S.: Conserving Land-Atmosphere Synthesis Suite (CLASS) v 1.1, , doi:10.25914/5c872258dc183,
984   2019.
985   Hobeichi, S.: Derived Optimal Linear Combination Evapotranspiration - DOLCE v2.1, ,
986   doi:10.25914/5f1664837ef06, 2020.

987 Hobeichi, S., Abramowitz, G., Evans, J. and Ukkola, A.: Derived Optimal Linear Combination
988 Evapotranspiration (DOLCE): A global gridded synthesis et estimate, Hydrol. Earth Syst. Sci., 22(2), 1317–
989 1336, doi:10.5194/hess-22-1317-2018, 2018.
990 Hobeichi, S., Abramowitz, G., Evans, J. and Beck, H. E.: Linear Optimal Runoff Aggregate (LORA): A global
991 gridded synthesis runoff product, Hydrol. Earth Syst. Sci., 23, 851–870, doi:10.5194/hess-23-851-2019,
992 2019.
993 Hobeichi, S., Abramowitz, G. and Evans, J. P.: Conserving Land – Atmosphere Synthesis Suite (CLASS), J.
994 Clim., 33, 1821–1844, doi:10.1175/JCLI-D-19-0036.1, 2020a.
995 Hobeichi, S., Abramowitz, G., Contractor, S. and Evans, J.: Evaluating precipitation datasets using surface
996 water and energy budget closure, J. Hydrometeorol., 989–1009, doi:10.1175/jhm-d-19-0255.1, 2020b.
997 Van Der Horst, S. V. J., Pitman, A. J., De Kauwe, M. G., Ukkola, A., Abramowitz, G. and Isaac, P.: How
998 representative are FLUXNET measurements of surface fluxes during temperature extremes?,
999 Biogeosciences, 16(8), 1829–1844, doi:10.5194/bg-16-1829-2019, 2019.
1000 Jiménez, C., Martens, B., Miralles, D. M., Fisher, J. B., Beck, H. E. and Fernández-prieto, D.: Exploring the
1001 merging of the global land evaporation WACMOS-ET products based on local tower measurements,
1002 Hydrol. Earth Syst. Sci, 22, 4513–4533, doi:https://doi.org/10.5194/hess-22-4513-2018, 2018.
1003 Jung, M., Reichstein, M., Ciais, P., Seneviratne, S. I., Sheffield, J., Goulden, M. L., Bonan, G., Cescatti, A.,
1004 Chen, J., De Jeu, R., Dolman, A. J., Eugster, W., Gerten, D., Gianelle, D., Gobron, N., Heinke, J., Kimball, J.,
1005 Law, B. E., Montagnani, L., Mu, Q., Mueller, B., Oleson, K., Papale, D., Richardson, A. D., Roupsard, O.,
1006 Running, S., Tomelleri, E., Viovy, N., Weber, U., Williams, C., Wood, E., Zaehle, S. and Zhang, K.: Recent
1007 decline in the global land evapotranspiration trend due to limited moisture supply, Nature, 467(7318),
1008 951–954, doi:10.1038/nature09396, 2010.
1009 Jung, M., Koirala, S., Weber, U., Ichii, K., Gans, F., Gustau-Camps-Valls, Papale, D., Schwalm, C.,
1010 Tramontana, G. and Reichstein, M.: The FLUXCOM ensemble of global land-atmosphere energy fluxes, ,
1011 6:74, 1–14, doi:10.1038/s41597-019-0076-8, 2019.
1012 Kanamitsu, M., Ebisuzaki, W., Woollen, J., Yang, S.-K., Hnilo, J. J., Fiorino, M. and Potter, G. L.: NCEP-DOE
1013 AMIP-II Reanalysis (R-2), Bull. Am. Meteorol. Soc., 83(11), 1631–1644, doi:10.1175/BAMS-83-11-1631,
1014 2002.
1015 Kendall, M. G.: Rank correlation methods., 1948.
1016 L'Ecuyer, T. S., Beaudoing, H. K., Rodell, M., Olson, W., Lin, B., Kato, S., Clayson, C. A., Wood, E.,
1017 Sheffield, J., Adler, R., Huffman, G., Bosilovich, M., Gu, G., Robertson, F., Houser, P. R., Chambers, D.,
1018 Famiglietti, J. S., Fetzer, E., Liu, W. T., Gao, X., Schlosser, C. A., Clark, E., Lettenmaier, D. P. and Hilburn,
1019 K.: The observed state of the energy budget in the early twenty-first century, J. Clim., 28(21), 8319–
1020 8346, doi:10.1175/JCLI-D-14-00556.1, 2015.
1021 Liang, X., Lettenmaier, D. P., Wood, E. F. and Burges, S. J.: A simple hydrologically based model of land
1022 surface water and energy fluxes for general circulation models, J. Geophys. Res. Atmos., 99(D7), 14415--
1023 14428, doi:10.1029/94JD00483, 1994.
1024 Lloyd, S.: Least squares quantization in PCM, IEEE Trans. Inf. theory, 28(2), 129–137, 1982.
1025 Long, D., Longuevergne, L. and Scanlon, B. R.: Uncertainty in evapotranspiration fromland
1026 surfacemodeling, remote sensing, and GRACE satellites, Water Resour. Res., 50(2), 1131–1151,
1027 doi:10.1002/2013WR014581.Received, 2014.
1028 Lorenz, C., Tourian, M. J., Devaraju, B., Sneeuw, N. and Kunstmann, H.: Basin-scale runoff prediction: An
1029 Ensemble Kalman Filter framework based on global hydrometeorological data sets, Water Resour. Res.,
1030 51, 8450–8475, doi:10.1002/2014WR016794, 2015.

1031   MacQueen, J.: Some methods for classification and analysis of multivariate observations, in Proceedings
1032   of the fifth Berkeley symposium on mathematical statistics and probability, vol. 1, pp. 281–297, Oakland,
1033   CA, USA., 1967.
1034   Mann, H. B.: Nonparametric tests against trend, Econom. J. Econom. Soc., 245–259, 1945.
1035   Marengo, J. A., Souza, C. M., Thonicke, K., Burton, C., Halladay, K., Betts, R. A., Alves, L. M. and Soares,
1036   W. R.: Changes in Climate and Land Use Over the Amazon Region: Current and Future Variability and
1037   Trends, Front. Earth Sci., 6, doi:10.3389/feart.2018.00228, 2018.
1038   Martens, B., Miralles, D., Lievens, H., Van Der Schalie, R., De Jeu, R., Fernández-Prieto, D. and Verhoest,
1039   N.: GLEAM v3: updated land evaporation and root-zone soil moisture datasets, Geophys. Res. Abstr. EGU
1040   Gen. Assem., 18, 2016–4253, 2016.
1041   Martens, B., Miralles, D. G., Lievens, H., Van Der Schalie, R., De Jeu, R. A. M., Fernández-Prieto, D., Beck,
1042   H. E., Dorigo, W. A. and Verhoest, N. E. C.: GLEAM v3: Satellite-based land evaporation and root-zone
1043   soil moisture, Geosci. Model Dev., 10(5), 1903–1925, doi:10.5194/gmd-10-1903-2017, 2017.
1044   McCabe, M. F., Ershadi, A., Jimenez, C., Miralles, D. G., Michel, D. and Wood, E. F.: The GEWEX LandFlux
1045   project: Evaluation of model evaporation using tower-based and globally gridded forcing data, Geosci.
1046   Model Dev., 9, 283–305, doi:10.5194/gmd-9-283-2016, 2016.
1047   Michel, D., Jiménez, C., Miralles, D. G., Jung, M., Hirschi, M., Ershadi, A., Martens, B., McCabe, M. F.,
1048   Fisher, J. B., Mu, Q., Seneviratne, S. I., Wood, E. F. and Fernández-Prieto, D.: The WACMOS-ET project
1049   Part 1: Tower-scale evaluation of four remote-sensing-based evapotranspiration algorithms, Hydrol.
1050   Earth Syst. Sci., 20(2), 803–822, doi:10.5194/hess-20-803-2016, 2016.
1051   Miralles, D. G., Holmes, T. R. H., De Jeu, R. A. M., Gash, J. H., Meesters, A. G. C. A. and Dolman, A. J.:
1052   Global land-surface evaporation estimated from satellite-based observations, Hydrol. Earth Syst. Sci.,
1053   15(2), 453–469, doi:10.5194/hess-15-453-2011, 2011a.
1054   Miralles, D. G., De Jeu, R. A. M., Gash, J. H., Holmes, T. R. H. and Dolman, A. J.: Magnitude and variability
1055   of land evaporation and its components at the global scale, Hydrol. Earth Syst. Sci., 15(3), 967–981,
1056   doi:10.5194/hess-15-967-2011, 2011b.
1057   Miralles, D. G., Van Den Berg, M. J., Gash, J. H., Parinussa, R. M., De Jeu, R. A. M., Beck, H. E., Holmes, T.
1058   R. H., Jiménez, C., Verhoest, N. E. C., Dorigo, W. A., Teuling, A. J., & Johannes Dolman, A. (2014). El Niño-
1059   La Niña cycle and recent trends in continental evaporation. In Nature Climate Change (Vol. 4, Issue 2, pp.
1060   122–126). https://doi.org/10.1038/nclimate2068.
1061   Montano, B. Q., Westerberg, I., Wetterhall, F., Hidalgo, H. G. and Halldin, S.: Characterising droughts in
1062   Central America with uncertain hydro-meteorological data, in 2015 AGU Fall Meeting, AGU., 2015.
1063   Mu, Q., Zhao, M. and Running, S. W.: Improvements to a MODIS global terrestrial evapotranspiration
1064   algorithm, Remote Sens. Environ., 115(8), 1781–1800, doi:10.1016/j.rse.2011.02.019, 2011.
1065   Mueller, B., Seneviratne, S. I., Jimenez, C., Corti, T., Hirschi, M., Balsamo, G., Ciais, P., Dirmeyer, P.,
1066   Fisher, J. B., Guo, Z., Jung, M., Maignan, F., McCabe, M. F., Reichle, R., Reichstein, M., Rodell, M.,
1067   Sheffield, J., Teuling, A. J., Wang, K., Wood, E. F. and Zhang, Y.: Evaluation of global observations-based
1068   evapotranspiration datasets and IPCC AR4 simulations, Geophys. Res. Lett., 38(6), 3–10,
1069   doi:10.1029/2010GL046230, 2011.
1070   Mueller, B., Hirschi, M., Jimenez, C., Ciais, P., Dirmeyer, P. A., Dolman, A. J., Fisher, J. B., Jung, M.,
1071   Ludwig, F., Maignan, F., Miralles, D. G., McCabe, M. F., Reichstein, M., Sheffield, J., Wang, K., Wood, E.
1072   F., Zhang, Y. and Seneviratne, S. I.: Benchmark products for land evapotranspiration: LandFlux-EVAL
1073   multi-data set synthesis, Hydrol. Earth Syst. Sci., 17, 3707–3720, doi:10.5194/hess-17-3707-2013, 2013.

1074   Munier, S., Aires, F., Schlaffer, S., Prigent, C., Papa, F., Maisongrande, P. and Pan, M.: Combining
1075   datasets of satellite retrieved products for basin-scale water balance study. Part II: Evaluation on the
1076   Mississippi Basin and closure correction model, J. Geophys. Res. Atmos., 119, 12,100-12,116,
1077   doi:10.1002/2014JD021953, 2014.
1078   Paca, V. H. da M., Espinoza-Dávalos, G. E., Hessels, T. M., Moreira, D. M., Comair, G. F. and Bastiaanssen,
1079   W. G. M.: The spatial variability of actual evapotranspiration across the Amazon River Basin based on
1080   remote sensing products validated with flux towers, Ecol. Process., 8(1), doi:10.1186/s13717-019-0158-
1081   8, 2019.
1082   Pan, S., Pan, N., Tian, H., Friedlingstein, P., Sitch, S., Shi, H., Arora, V. K., Haverd, V., Jain, A. K., Kato, E.,
1083   Lienert, S., Lombardozzi, D., Ottle, C., Poulter, B. and Zaehle, S.: Evaluation of global terrestrial
1084   evapotranspiration by state-of-the-art approaches in remote sensing, machine learning, and land
1085   surface models, Hydrol. Earth Syst. Sci., 24, 1485–1509, doi:10.5194/hess-24-1485-2020, 2020.
1086   Rodell, M., Beaudoing, H. K., L'Ecuyer, T. S., Olson, W. S., Famiglietti, J. S., Houser, P. R., Adler, R.,
1087   Bosilovich, M. G., Clayson, C. A., Chambers, D., Clark, E., Fetzer, E. J., Gao, X., Gu, G., Hilburn, K.,
1088   Huffman, G. J., Lettenmaier, D. P., Liu, W. T., Robertson, F. R., Schlosser, C. A., Sheffield, J. and Wood, E.
1089   F.: The observed state of the water cycle in the early twenty-first century, J. Clim., 28, 8289–8318,
1090   doi:10.1175/JCLI-D-14-00555.1, 2015.
1091   Sahoo, A. K., Pan, M., Troy, T. J., Vinukollu, R. K., Sheffield, J. and Wood, E. F.: Reconciling the global
1092   terrestrial water budget using satellite remote sensing, Remote Sens. Environ., 115, 1850–1865,
1093   doi:10.1016/j.rse.2011.03.009, 2011.
1094   Sen, P. K.: Estimates of the regression coefficient based on Kendall's tau, J. Am. Stat. Assoc., 63(324),
1095   1379–1389, 1968.
1096   Sharma, A., Wasko, C. and Lettenmaier, D. P.: If precipitation extremes are increasing, why aren't
1097   floods?, Water Resour. Res., 54(11), 8545–8551, 2018.
1098   Sheffield, J., Wood, E. F. and Roderick, M. L.: Little change in global drought over the past 60 years,
1099   Nature, 491(7424), 435–438, doi:10.1038/nature11575, 2012.
1100   Stackhouse Jr, P. W., Gupta, S. K., Cox, S. J., Zhang, T., Mikovitz, J. C. and Hinkelman, L. M.: 24.5-Year
1101   surface radiation budget data set released, Glob. Energy Water Cycle Exp. News, 21(1), 1–20, 2011.
1102   Stephens, G. L., Li, J., Wild, M., Clayson, C. A., Loeb, N., Kato, S., L'Ecuyer, T., Stackhouse, P. W., Lebsock,
1103   M. and Andrews, T.: An update on Earth's energy balance in light of the latest global observations, Nat.
1104   Geosci., 5, 691–696, doi:10.1038/ngeo1580, 2012.
1105   Su, Z.: The Surface Energy Balance System (SEBS) for estimation of turbulent heat fluxes, Hydrol. Earth
1106   Syst. Sci., 6(1), 85–100, doi:10.5194/hess-6-85-2002, 2002.
1107   Teuling, A. J.: A hot future for European droughts, Nat. Clim. Chang., 8(5), 364–365, 2018.
1108   Teuling, A. J., Van Loon, A. F., Seneviratne, S. I., Lehner, I., Aubinet, M., Heinesch, B., Bernhofer, C.,
1109   Grünwald, T., Prasse, H. and Spank, U.: Evapotranspiration amplifies European summer drought,
1110   Geophys. Res. Lett., 40(10), 2071–2075, 2013.
1111   Ukkola, A. M., Pitman, A. J., Donat, M. G., De Kauwe, M. G. and Angélil, O.: Evaluating the Contribution
1112   of Land-Atmosphere Coupling to Heat Extremes in CMIP5 Models, Geophys. Res. Lett., 45(17), 9003–
1113   9012, doi:10.1029/2018GL079102, 2018.
1114   Vinukollu, R. K., Wood, E. F., Ferguson, C. R. and Fisher, J. B.: Global estimates of evapotranspiration for
1115   climate studies using multi-sensor remote sensing data: Evaluation of three process-based approaches,
1116   Remote Sens. Environ., 115, 801–823, doi:10.1016/j.rse.2010.11.006, 2011.

1117   Wan, Z., Zhang, K., Xue, X., Hong, Z., Hong, Y. and J. Gourley, J.: Water balance-based actual

1118   evapotranspiration reconstruction fromground and satellite observations over the conterminous United

1119   States Zhanming, Water Resour. Res., 51, 6485–6499, doi:10.1002/2015WR017311, 2015.

1120   Zhang, K., Kimball, J. S., Nemani, R. R. and Running, S. W.: A continuous satellite-derived global record of

1121   land surface evapotranspiration from 1983 to 2006, Water Resour. Res., 46(9),

1122   doi:10.1029/2009WR008800, 2010.

1123   Zhang, K., Kimball, J. S., Nemani, R. R., Running, S. W., Hong, Y., Gourley, J. J. and Yu, Z.: Vegetation

1124   Greening and Climate Change Promote Multidecadal Rises of Global Land Evapotranspiration, Sci. Rep.,

1125   5(June), 1–9, doi:10.1038/srep15956, 2015.

1126   Zhang, Y., Peña-Arancibia, J. L., McVicar, T. R., Chiew, F. H. S., Vaze, J., Liu, C., Lu, X., Zheng, H., Wang, Y.,

1127   Liu, Y. Y., Miralles, D. G. and Pan, M.: Multi-decadal trends in global terrestrial evapotranspiration and its

1128   components, Sci. Rep., 6, 19124, doi:10.1038/srep19124, 2016.

1129   Zhang, Y., Pan, M., Sheffield, J., Siemann, A. L., Fisher, C. K., Liang, M., Beck, H. E., Wanders, N.,

1130   Maccracken, R. F., Houser, P. R., Zhou, T., Lettenmaier, D. P., Pinker, R. T., Bytheway, J., Kummerow, C.

1131   D. and Wood, E. F.: A Climate Data Record (CDR) for the global terrestrial water, Earth Syst. Sci, 22(1),

1132   241–263, doi:10.5194/hess-22-241-2018, 2018.

1133

1134