Journal of Hydrology 624 (2023) 129971

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Research papers

The divergence of energy- and water-balance evapotranspiration estimates in humid regions

Lilin Zhang *, Michael Marshall, Anton Vrieling, Andrew Nelson

University of Twente, Faculty of Geo-Information Science and Earth Observation (ITC), 7500 AE Enschede, the Netherlands

ARTICLE INFO

This manuscript was handled by Emmanouil Anagnostou, Editor-in-Chief

Keywords:
Catchment water balance
Spatiotemporal variability
Budyko framework
Terrestrial water storage anomaly
Excess precipitation
Energy constraints

ABSTRACT

Evapotranspiration (ET) calculated as the residual of catchment water balance (ET_{WB}) has often been used as a benchmark to evaluate satellite-based ET retrievals that use the energy-balance approach (ET_{EB}). However, errors from water balance components will accrue in ET_{WB}, leading to considerable disparities with ET_{EB}. In this study, we set out to investigate whether ET_{EB} from multiple sources (MOD16, GLEAM, PT-JPL, and PT-hybrid) can capture the spatiotemporal variability of ET_{WB} across 53 catchments in central-western Europe with a humid climate. Using ET retrievals from the Budyko framework that accounts for the control of energy demand on water supply and upscaled ET from FLUXCOM as references, we explored the causes of discrepancies between ET_{WB} and ET_{EB} at long-term, annual, and monthly scales. We found that (1) ET_{EB} significantly diverged from ET_{WB} at the mean annual scale ($r = 0.35$), particularly for energy-limited catchments, but Budyko-simulated ET considering energy limit correlated well with ET_{EB} ($r > 0.86$); (2) neither ET_{EB} nor upscaled ET can reproduce annual ET_{WB} time series ($r < 0.40$), and the closure errors in water budgets closely follow excess precipitation beyond energy demand; (3) monthly ET_{WB} exhibited better correspondences with ET_{EB} ($r = 0.73$), presumably because of similarity in seasonal patterns. Our results demonstrate that errors from precipitation and terrestrial water storage anomalies introduce large uncertainties in ET_{WB}, thereby complicating water balance validation in humid regions across multiple timesteps. To improve the application of ET_{WB} for benchmarking ET_{EB} in humid regions, high-quality input data should be used or – like the Budyko framework – energy constraints should be considered.

1. Introduction

As water supplies become increasingly limited, accurate quantification of the Earth's surface water resources has become crucial for balancing terrestrial water demand and water availability (D'Odorico et al., 2018). Terrestrial evapotranspiration (ET), the water leaving the Earth's surface and entering the atmosphere, is the second largest component of the water cycle after precipitation and plays a vital role in global hydrological and energy cycles (Miralles et al., 2011). ET cannot be directly measured by remote sensing but can be derived from satellite-based land surface variables by using the energy-balance approach, referred to as ET_{EB} (Glenn et al., 2007). Various ET_{EB} methods with different scopes and complexities have been developed, such as ALEXI (Anderson et al., 2007), PT-JPL (Fisher et al., 2008), MOD16 (Mu et al., 2011), GLEAM (Miralles et al., 2011), and LSA-SAF ET (Ghilain et al., 2011). These ET_{EB} methods rely on satellite-driven indicators to simulate land–atmosphere interactions and monitor large-scale ET variability under different weather conditions (Biggs et al., 2015). Nevertheless, there are substantial differences among these ET_{EB} products due to varying parameters and inputs, leading to inconsistent ET trends at continental and catchment scales depending on the chosen product (Badgley et al., 2015; Hu et al., 2015; Zhu et al., 2022). Therefore, before ET_{EB} estimates can be used reliably in hydrological and agricultural applications, they should be rigorously evaluated against reference observations.

ET_{EB} estimates are usually evaluated against eddy-covariance flux tower measurements at the field scale (Wang and Dickinson, 2012). FLUXNET is one of the flux tower networks that facilitates ET validation across numerous sites with a diversity of vegetation types (Michel et al., 2016). However, the representativeness of ET measurements is limited by the uneven density of flux towers, which concentrate on Europe and US (McCabe and Wood, 2006; Pastorello et al., 2020). Other observational ET approaches such as Bowen ratio systems (Bowen, 1926), weighing lysimeters (Holmes, 1984), surface renewal (Kyaw Tha Paw

* Corresponding author.
E-mail address: l.zhang-2@utwente.nl (L. Zhang).

https://doi.org/10.1016/j.jhydrol.2023.129971
Received 19 May 2023; Received in revised form 12 July 2023; Accepted 16 July 2023
Available online 20 July 2023

et al., 1995) and large aperture scintillometers (Meijninger et al., 2002) can serve as the field-level references, but are even more limited over space and time. Due to the lack of representativeness of these datasets, regional evaluation of ET_{EB} estimates remains a challenge.

Efforts have been made to scale ET from available *in situ* measurements to regional references by using machine learning methods (Jung et al., 2009; Yao et al., 2017). For example, the Multi-Tree Ensemble approach was first applied to scale up *in situ* measured ET from the global FLUXNET network with remote sensing and gridded meteorological data (Jung et al., 2010). The machine learning-based upscaling method's validity for generating energy flux estimates and reproducing ET spatial patterns is well-documented (Jung et al., 2019; Jung et al., 2011). As a result, the upscaled ET as well as the extended FLUXCOM product has been commonly used as a regional proxy for FLUXNET observations to validate ET_{EB} (Brust et al., 2021; Jiang and Ryu, 2016; Velpuri et al., 2013). The reliability of the scaled reference data highly depends on how comprehensively the training datasets describe global ecosystem behaviours (Pan et al., 2020). Although machine learning methods improve the representativeness of flux tower data in spatially heterogeneous regions, their accuracy may be low outside the geographic and climate range of the training data.

An alternative method for determining regional ET is the water-balance approach that calculates water balance ET (ET_{WB}) by subtracting discharge from precipitation over a long period of time, when terrestrial water storage anomaly (TWSA) is assumed to be negligible (Hobbins et al., 2001). In reality, the fundamental assumption that TWSA can be neglected may be problematic for some basins because of snow melt, lake-level change, and human factors such as reservoir regulation and irrigation (Han et al., 2020; Rodell et al., 2018). The Gravity Recovery and Climate Experiment (GRACE) satellites currently provide a unique way to measure TWSA, but their processed water storage data are available at coarse spatial resolution (0.25–1°) (Save et al., 2016; Wiese et al., 2016). As more GRACE data have accumulated and the data processing strategies have improved, recent studies have demonstrated the potential of GRACE data to provide water storage information at finer spatial scales (Scanlon et al., 2016; Senay et al., 2016; Zhang et al., 2018). For example, Pascolini-Campbell et al. (2021) found that the modelled TWSA from GRACE and the GRACE Follow-On mission can detect agricultural water use in catchments smaller than 10,000 km^2, presumably because terrestrial water storage was validated to be spatially homogenous within the region. Given that it is hard to make a compromise between using coarse-resolution GRACE data and assuming TWSA as negligible, both ET_{WB} considering GRACE-based TWSA (Bai and Liu, 2018; Pascolini-Campbell et al., 2020) and ET_{WB} ignoring TWSA (Marshall et al., 2012; Yin et al., 2020) have been extensively utilised to validate ET_{EB} in the past decades.

Although ET_{WB} estimates have been increasingly utilised to validate ET_{EB}, these estimates have uncertainties because of errors in precipitation, runoff and TWSA, which makes the use of ET_{WB} as benchmark problematic. Numerous studies have revealed significant divergences between the magnitude of ET_{EB} and ET_{WB}, with imbalance errors as high as 25% of mean annual precipitation (Sahoo et al., 2011; Zhang et al., 2012; Zhang et al., 2018). Moreover, closure errors of water budgets followed distinct climatic gradients, with humid regions exhibiting lager disparities than arid regions (Liu et al., 2016). With respect to consistency, Bai and Liu (2018) found that independent ET products had an average correlation of 0.33 with ET_{WB} for 22 catchments at the annual scale, whereas higher correlations ($r = 0.84$) were observed when compared with monthly flux measurements. The inconsistency between ET_{WB} and ET_{EB} was also noted by Pang et al. (2021), who indicated that conclusions about the performance of ET_{EB} products based on ET_{WB} may be biased. Overall, due to differences in timesteps, study region, and forcing data, the relationship between ET_{WB} and ET_{EB} in terms of the consistency and magnitude differed greatly across previous studies (Ma et al., 2021; Yin et al., 2020). A comprehensive understanding of the mismatch between ET_{WB} and ET_{EB} is crucial for accurately interpreting water balance validation, especially for humid regions.

Efforts have been made to account for the disparities between ET_{WB} and ET_{EB} by resolving water cycle imbalances (Lehmann et al., 2022). Previous studies attempted to reduce precipitation errors by using an ensemble of precipitation datasets, but this did not result in effectively closing water budgets (Pascolini-Campbell et al., 2020; Ruhoff et al., 2022). A widely-accepted explanation is that the disagreement of ET_{EB}-ET_{WB} is a result of errors in ET_{EB} products, such as model structure constraints and forcing dataset uncertainties (Liu et al., 2016). However, comparisons between ET_{EB} products and other independent ET sources suggest that the inconsistency of ET_{WB}-ET_{EB} in humid regions might be equally influenced by uncertainties in water balance components rather than solely within ET_{EB} datasets themselves (Carter et al., 2018; Han et al., 2015; Zhang et al., 2012). To investigate this, Pan et al. (2017) incorporated ET_{WB} using *in situ* TWSA as reference, and found large discrepancies between ET_{WB} and *in situ* TWSA-based ET_{WB} at both annual and monthly timesteps. Li et al. (2019) compared ET_{WB} with atmospheric-inferred ET using the atmospheric water-balance approach and concluded that atmospheric-inferred ET outperformed ET_{WB} as a benchmark in runoff-dominant catchments. Although these studies were conducted for a limited number of catchments in China, they confirmed that uncertainties involved in the calculation of ET_{WB} also hold responsible for the divergence of energy- and water-balance ET estimates. We contend that the non-closure issue of water budgets has not been sufficiently scrutinised and there is scope to investigate in more depth the reasons for the divergences of ET_{WB}-ET_{EB} across different timesteps.

Central-western Europe, characterized by a humid climate and a dense network for precipitation and discharge monitoring, is a suitable test case for comparing ET_{WB} and ET_{EB} in humid regions. To give insight into the divergence of ET_{WB}-ET_{EB}, we employed the Budyko framework, a robust approach that efficiently describes the relationship between long-term ET and terrestrial energy- and water- balances at catchment scales (Budyko, 1974; Fu, 1981; Zhang et al., 2010). Unlike the water-balance approach that calculates ET as the residual of water balance equation, the Budyko framework estimates ET through hydrological partitioning and has been widely used for benchmarking ET estimates and calibrating ET algorithms (Kim et al., 2022; Koppa and Gebremichael, 2017; Zhang et al., 2010). Additionally, considering that machine learning-based upscaling methods perform better in areas with more ground truth on energy flux, such as US and Europe, we also incorporated upscaled ET from FLUXCOM product for interpreting water balance validation at annual and sub-annual scales. Our study had three main objectives: (1) to assess the impacts of coarse-resolution GRACE data on long-term and annual water balances for small-scale catchments; (2) to evaluate the consistency between ET_{WB} and ET_{EB}, where ET_{WB} serves as an independent data source to be compared with multisource ET_{EB} (GLEAM, MOD16, PT-JPL, and PT-hybrid); and (3) to investigate possible reasons for the varying divergence of ET_{WB}-ET_{EB} across different timesteps. In contrast to previous studies that treated ET_{WB} as benchmark without accounting for uncertainties within the water balance calculation, our study aimed to enhance the understanding of closure errors in water budgets at different time scales and shed light on the limitations of ET_{WB} in energy-limited catchments.

2. Materials and methods

Multiple datasets were used to comprehensively evaluate the differences between ET_{WB} and ET_{EB} across 53 catchments in central-western Europe. Table 1 gives an overview of the gridded datasets plus the basin-wide ET derived from the Budyko framework. To facilitate intercomparison, all gridded datasets except coarse-resolution GRACE data were resampled to a common 5 km spatial resolution based on nearest-neighbor interpolation and were then aggregated to a monthly time step. Fig. 1 displays the workflow of water balance assessment in this study. We used four different ET_{EB} models: PT-JPL,

Table 1
Overview of gridded datasets used in this study together with the basin-wide ET derived from the Budyko framework.

Variable	Data	Time range	Spatial resolution	Temporal resolution	Source
ET_{EB}	PT-JPL	2003–2020	1 km	Monthly	Fisher et al. (2008)
	PT-JPL (hybrid)	2003–2020	1 km	Monthly	Zhang et al. (2021)
	MOD16	2003–2020	500 m	8-day	Mu et al. (2011)
	GLEAM	2003–2020	25 km	Daily	Martens et al. (2017)
Precipitation	ERA5-Land	2003–2020	10 km	Monthly	Muñoz Sabater (2019)
	WorldClim	2003–2018	5 km	Monthly	Fick and Hijmans (2017)
	E-OBS	2003–2020	10 km	Daily	Cornes et al. (2018)
TWSA	CSR-GRACE	2003–2020	25 km	Monthly	Save et al. (2016)
	GFZ-GRACE	2003–2020	100 km	Monthly	Boergens et al. (2019)
	JPL-GRACE	2003–2020	50 km	Monthly	Wiese et al. (2016)
ET references	FLUXCOM	2003–2015	1 km	8-day	Jung et al. (2019)
	ET from the Budyko framework	2003–2020	Basin wide	Long-term	This study

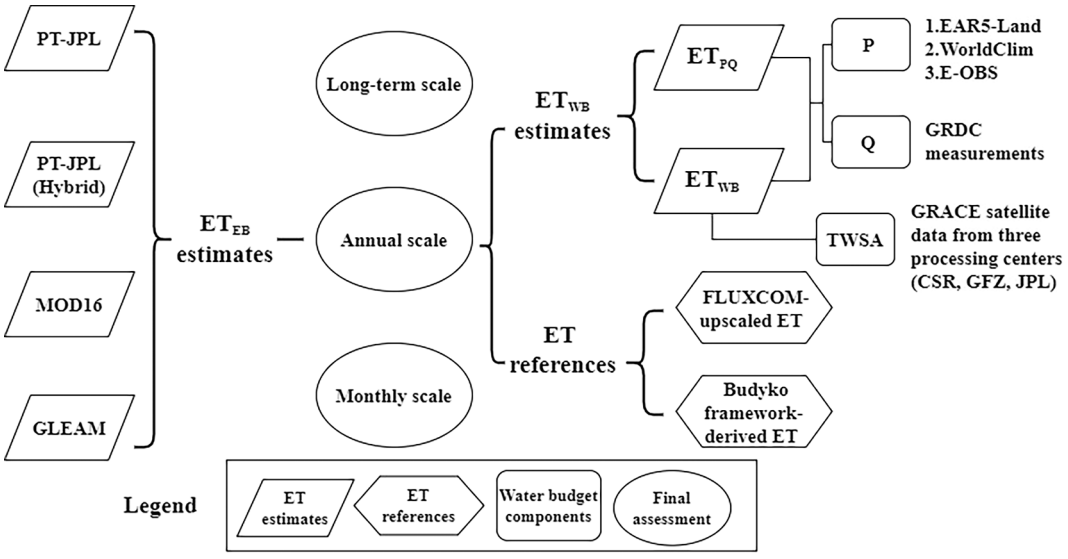

Fig. 1. Workflow describing the water balance assessment in this study.

PT-hybrid, MOD16, and GLEAM for water balance comparison, which have been available since 2003. The ET references in this study included upscaled ET from FLUXCOM product and ET estimates from the Budyko framework.

2.1. ET_{EB} estimates

The PT-JPL model is based on the Priestley-Taylor algorithm that calculates ET as a fraction of potential evapotranspiration (PET) by using a series of ecophysiological constraints (Fisher et al., 2008). The PT-JPL method's good accuracy has been validated in various studies (Ershadi et al., 2014; Fisher et al., 2020; Vinukollu et al., 2011), which has been employed as the primary ET estimation method for NASA's ECOSTRESS mission. The PT-JPL ET estimates in our study were derived from MODIS reflectance data and ERA5-Land meteorological reanalysis data that provide relative humidity, air temperature, and surface net radiation (Rn) calculated as the sum of surface net solar radiation and surface net thermal radiation. For more details of the generation of gridded PT-JPL ET as well as the PT-hybrid ET at a monthly timestep, see Zhang et al. (2021).

The PT-hybrid method is a modified version of the PT-JPL method, which uses optical shortwave infrared (SWIR)-based spectral indices and microwave soil moisture to parameterise the soil moisture constraint (fsm) for cropland and grassland, respectively. For other landscapes such as forest areas, fsm is calculated using the original parameterisation based on the relative humidity/vapour pressure deficit defined in Fisher et al. (2008). Additionally, the MODIS land cover product (MCD12Q1) at 500 m resolution was utilised to determine cropland and grassland landscapes. ET estimates from the PT-JPL and PT-hybrid methods have been validated over 31 FLUXNET sites with an average correlation coefficient of 0.86 (Zhang et al., 2021).

The MODIS ET algorithm, the most widely used ET model at global scale, is based on a modified Penman-Monteith equation in which the aerodynamic resistance is calculated as a parallel resistance to convective and radiative heat transfer, and the canopy conductance for plant transpiration is calculated using the Leaf Area Index (Mu et al., 2007; Mu et al., 2011). The MOD16 ET product is derived from a series of MODIS datasets, including Leaf Area Index, land cover and albedo, and the Modern-Era Retrospective Analysis for Research and Applications (MERRA) meteorological reanalysis data. After calibration with AmeriFlux tower observations, the MOD16 provides global ET estimates at 8-day timesteps (Mu et al., 2011).

The GLEAM ET method is based on a modified Priestley-Taylor algorithm in which PET estimates have been converted into actual

evaporation by a stress factor based on microwave retrievals of vegetation optical depth and rootzone soil moisture (Miralles et al., 2011). The GLEAM v3b dataset adopted in our study was driven solely by remote sensing derived products, including data from the Clouds and Earth's Radiant Energy System (CERES), AIRS (Atmospheric Infra-Red Sounder), SMOS (Soil Moisture and Ocean Salinity), and ESA-CCI (European Space Agency—Climate Change Initiative). The GLEAM ET dataset has been validated as having comparable accuracy to PT-JPL using surface water balances from 837 globally distributed catchments (Miralles et al., 2016).

2.2. ET_{WB} estimates derived from water balance equation

ET_{WB} is calculated as the residual of terrestrial water balance at the basin scale by assuming no net groundwater flow across the boundary of the river basin, following:

$$ET_{WB} = P - Q - \Delta S \quad (1)$$

where P is precipitation, Q is river discharge (outflow minus inflow), and ΔS is the change in terrestrial water storage, which includes changes in groundwater and surface water storage (Wan et al., 2015). The monthly ΔS can be calculated from GRACE-based TWSA by differencing the preceding and following months and dividing by 2 months (Pascolini-Campbell et al., 2020). With respect to the trade-off between ignoring ΔS and considering GRACE-based TWSA for small-scale catchments, ET_{WB} calculated as P minus Q (hereafter ET_{PQ}) was included as a comparison at long-term and annual scales.

Precipitation data were obtained from three published precipitation datasets including the ERA5-Land reanalysis dataset, the WorldClim v2.1 dataset, as well as the E-OBS dataset that is purely based on *in situ* precipitation measurements. The ERA5-Land dataset produced by the European Centre for Medium-Range Weather Forecasts (ECMWF) uses the most recent Earth system and data assimilation methods and was shown to be an improved atmospheric reanalysis than the former ERA-Interim dataset (Albergel et al., 2018; He et al., 2021). The WorldClim v2.1 is a global climate dataset that combines weather station data with satellite covariates and has been widely used in ecology, biodiversity and agricultural studies (Fick and Hijmans, 2017). E-OBS is a new ensemble version of gauge-observation dataset available for Europe-wide, which has been commonly used for climate monitoring and model validation (Cornes et al., 2018). To get a more robust estimate of catchment precipitation, we extracted an ensemble mean of precipitation from these three data sources. Following Pascolini-Campbell et al. (2020), the standard deviation of the precipitation time series from all three sources was used to quantity P errors within the calculation of ET_{WB}.

River discharge data were collected from the Global Runoff Data Centre (GRDC), which is the world's largest and mostly extensively-used runoff dataset (https://www.bafg.de/GRDC). GRDC also provides watershed boundaries for around 7000 GRDC stations, although watershed polygons are missing for some catchments (GRDC, 2011). We selected catchments distributed over Europe, based on the following criteria: (a) having a drainage area larger than 1,000 km^2 with available watershed boundaries from GRDC; (b) having at least five years of continuous runoff data during 2003–2020; (c) having a runoff-rainfall coefficient less than 0.5 to exclude runoff-dominant catchments (e.g., mountainous catchments with high P and low ET), which is calculated as the ratio of long-term average runoff to precipitation. The selection process resulted in 53 catchments (See Fig. 2) and following Zhang et al. (2012), we split them into two groups: 25 large-area catchments (5000 to 150000 km^2), and 28 small-area catchments (1000 to 5000 km^2). See Table 2 for more details on the selected catchments. As suggested by Sauer and Meyer (1992), *in situ* measurements of discharge have the least uncertainty in water balance components and most discharge errors range from 3% to 6%. In this study, we set observational errors in Q as 5% of catchment discharge for long-term analysis, which is consistent with previous studies (Castle et al., 2016; Li et al., 2019).

The terrestrial water storage data were retrieved from the GRACE satellites launched in March 2002 and the GRACE Follow-On mission launched in May 2018, which monitor global gravity to make monthly

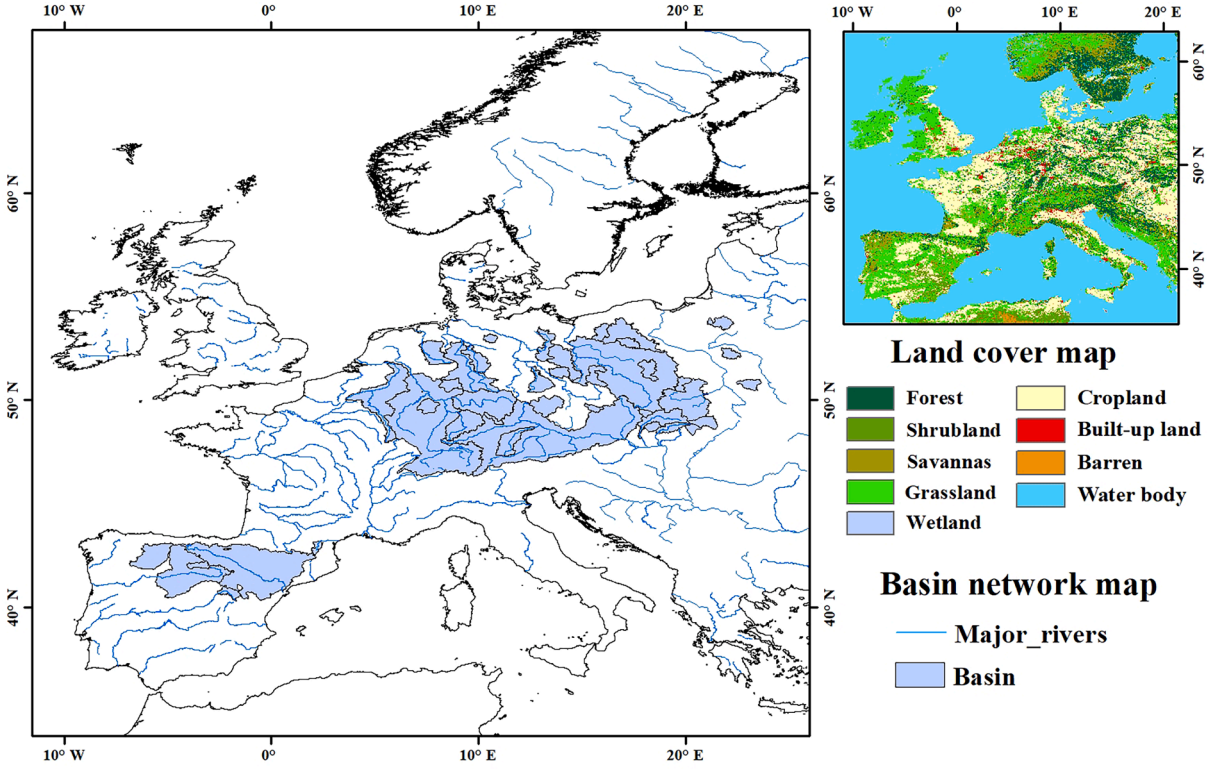

Fig. 2. Location of basins and land cover type of study region (2020).

Table 2
The 53 river basins that were part of the study with main characteristics. (P = mean annual precipitation; PET = mean annual potential evapotranspiration; AI = aridity index; Temporal availability is determined by runoff data availability from GRDC, Latitude and longitude denote the geographic location of the hydrological station where the runoff of the basin is measured).

Basin Name	River	Lat	Lon	Drainage area (km^2)	Elev. (m)	Temporal Availability	P (mm)	PET (mm)	AI
Affoldern	Eder	51.17	9.09	1434	193.19	2003–2018	942	668	0.71
Bad Dueben	Vereinigte Mulde	51.59	12.58	6164	81.50	2003–2019	905	645	0.71
Bad Liebenwerda	Schwarze Elster	51.52	13.4	3078	83.91	2003–2019	749	662	0.88
Berlin Muehlendamm	Spree	52.46	13.86	9506	28.01	2003–2019	713	664	0.93
Beroun	Berounka	49.96	14.09	8296	213.41	2003–2018	735	695	0.95
Bienenbuettel	Ilmenau	53.15	10.46	1480	14.43	2003–2019	755	624	0.83
Boguslaw	Prosna	51.9	17.95	4344	87.87	2003–2020	663	631	0.95
Bratislava	Danube	48.14	17.11	131,023	128.00	2003–2017	1076	705	0.66
Breclav-Ladna	Thaya	48.81	16.85	11,931	157.38	2003–2018	740	694	0.94
Brehy	Hron	48.41	18.65	3856	195.00	2003–2017	952	719	0.76
Breto	Esla	41.87	−5.76	14,354	691.00	2003–2011	743	962	1.30
Carcassonne	Aude	43.21	2.36	1825	155.10	2003–2012	977	989	1.01
Chmelnica	Poprad	49.29	20.73	1242	507.00	2003–2017	931	640	0.69
Cochem	Moselle	50.14	7.17	27,125	77.03	2003–2019	925	720	0.78
Doerverden	Weser	52.85	9.21	22,141	7.99	2003–2018	831	651	0.78
Drawiny	Drawa	52.89	15.98	3282	29.79	2003–2020	721	642	0.89
Duesseldorf	Rhine	51.23	6.77	147,470	24.48	2003–2019	1047	716	0.68
Eichstaett	Altmuhl	48.89	11.2	1405	382.19	2003–2019	838	724	0.86
Fraga	Cinca	41.52	0.35	9621	100.00	2003–2017	823	1043	1.27
Frankfurt Osthafen	Main	50.11	8.71	24,764	90.64	2003–2019	850	706	0.83
Goerlitz	Neisse	51.16	14.99	1617	175.63	2003–2019	894	637	0.71
Gozdowice	Oder	52.76	14.31	109,782	2.98	2003–2020	708	643	0.91
Grolsheim	Nahe	49.91	7.91	4006	85.00	2003–2016	776	716	0.92
Hamoir	Ourthe	50.44	5.53	1594	109.90	2003–2012	1019	673	0.66
Ketzin	Havel	52.48	12.85	15,472	28.40	2003–2019	704	661	0.94
Kowanowko	Welna	52.67	16.84	2764	51.24	2003–2020	649	617	0.95
Krasnystaw	Wieprz	50.99	21.18	2987	173.86	2003–2020	755	639	0.85
Landau	Isar	48.67	12.69	8807	333.65	2003–2019	1193	720	0.60
Lenartovce	Sajo	48.3	20.31	1803	150.00	2003–2017	852	729	0.86
Letzter Heller	Werra	51.41	9.71	5466	117.40	2003–2018	818	666	0.81
Leun Neu	Lahn	50.55	8.36	3579	134.99	2003–2019	834	680	0.82
Lith	Maas	51.82	5.42	28,886	5.00	2003–2018	922	682	0.74
Lochow	Liwiec	52.51	21.68	2419	94.91	2003–2020	713	609	0.85
Moravicany	Morava	49.76	16.98	1566	244.99	2003–2018	893	640	0.72
Nitrianska Streda	Nitra	48.52	18.17	2088	158.00	2003–2017	837	722	0.86
Nowe Drezdenko	Notec	52.85	13.84	16,071	24.21	2003–2020	678	624	0.92
Peral De Arlanza	Arlanza	42.08	−4.07	2417	766.00	2003–2017	715	967	1.35
Przedborz	Pilica	51.09	19.88	2567	187.22	2003–2020	748	648	0.87
Ptaki	Pisa	53.39	21.79	3452	104.77	2003–2020	712	606	0.85
Rockenau Ska	Necker	49.44	9.01	12,620	119.71	2003–2019	976	731	0.75
Sala	Vah	48.16	17.88	10,442	109.00	2003–2017	997	677	0.68
Schwarmstedt	Leine	52.68	9.6	6418	21.00	2003–2018	826	633	0.77
Seros	Segre	41.45	0.42	12,917	85.00	2003–2017	778	1014	1.30
Stein	Kocher	49.26	9.29	1926	154.14	2003–2019	913	724	0.79
Szczucin	Vistula	50.33	21.08	23,816	159.06	2003–2020	891	650	0.73
Teplice	Becva	49.53	17.75	1261	243.11	2003–2018	909	704	0.77
Tore	Douro	41.52	−5.41	41,924	637.00	2003–2017	584	948	1.62
Tortosa	Ebro	40.81	0.52	84,016	25.00	2003–2017	691	1002	1.45
Trillo	Tagus	40.7	−2.59	3259	727.00	2003–2017	593	1074	1.81
Versen Wehrdurchstich Gesamt	EMS	52.74	7.24	8456	6.71	2003–2016	836	623	0.75
Zagan	Bobr	51.62	15.32	4359	91.85	2003–2020	840	644	0.77
Zdana	Hornad	48.6	21.34	4247	167.00	2003–2017	827	705	0.85
Zruc Nad Sazavou	Sazava	49.74	15.1	1492	323.13	2003–2020	831	669	0.81

anomaly estimate. The TWSA retrieved from gravity data is expected to capture water fluxes from both natural dynamics and human activities. We obtained the most recent release (RL06) of GRACE Mascon data from the Center for Space Research (CSR) of the University of Texas at Austin and the Jet Propulsion Laboratory (JPL), and the Level-3 GRACE data provided by the Deutsches GeoForschungsZentrum Potsdam (GFZ). Previous studies found that the processed GRACE data can retrieve the true amplitudes of water storage in a concentrated area that is smaller than the original spatial resolution of datasets, presumably because aquifers are strongly connected in space and changes in storage water will gradually affect the surroundings (Crow et al., 2017; Wang et al., 2011). Consequently, in our study, ET_{WB} with and without considering TWSA are included to investigate the impacts of coarse-resolution GRACE data on water balances for small-scale catchments. Moreover, an ensemble mean (a simple arithmetic mean of JPL, CSR, and GFZ) was validated to be effective in reducing the noise in the gravity field products (Sakumura et al., 2014). To minimize the uncertainties of TWSA, we first resampled both JPL- and GFZ- GRACE data to the resolution of CSR-GRACE data at 25 km and then extracted a catchment average value of TWSA from the three datasets. Following Li et al. (2019), the total uncertainty in GRACE-based TWSA was calculated as the standard deviation of TWSA from three different sources.

2.3. ET references

2.3.1. Upscaled ET

The latest FLUXCOM data were generated by multiple machine-learning algorithms that established the relationships between land surface variables and energy fluxes (Jung et al., 2019). *In situ* ET measurements at the FLUXNET sites with geospatial information retrieved from the remote sensing and surface meteorological observations were integrated to yield a gridded ET product. The upscaled ET dataset comprises two products with different configurations: (1) RS Setup: nine machine learning methods based solely on continuous time series of MODIS remote sensing data from 2001 to 2015, and (2) RS + METEO Setup: three machine learning techniques based on daily meteorological data and mean seasonal cycles of satellite data from 2001 to 2013. The RS setup has the advantage of not requiring climate forcing datasets as input, which are themselves subject to uncertainty (Jung et al., 2019). Therefore, we only used the RS version of the FLUXCOM ET products.

2.3.2. The Budyko framework

The Budyko framework is a conceptual approach that estimates mean annual ET as a partition of available water, which is predominantly controlled by both water supply (expressed in terms of precipitation) and energy demand (often denoted by PET) (Budyko, 1974). Compared with other land surface models, the Budyko framework provides a simple but powerful tool to describe the response of ET to environmental change on multiyear scale (Bai et al., 2020). Due to the framework's effectiveness to predict the catchment energy and water fluxes, this approach has recently seen a renaissance in hydrological research (Kim et al., 2022; Wang et al., 2016; Zhang et al., 2010). Based on the Budyko framework, Fu (1981) proposed a widely used Budyko-type equation, referred to as Fu's equation, to estimate mean annual ET (Zhang et al., 2004). In Fu's equation, the evaporative index is expressed as a function of the aridity index (AI):

$$\frac{ET}{P} = 1 + \frac{PET}{P} - \left[1 + \left(\frac{PET}{P}\right)^{w}\right]^{1/w} \tag{2}$$

$$PET = \alpha \frac{\Delta}{\Delta + \gamma}(Rn - G) \tag{3}$$

The w in Fu's equation is an empirical parameter ranging from 1 to ∞ and reflects the impact of factors such as land surface characteristics and climate seasonality on water and energy balances. A higher w corresponds to an increase in ET efficiency, which means higher ET and lower runoff for a given precipitation and PET. To calibrate the single parameter w, ET is derived from multi-year historical records of precipitation and discharge data. For the calculation of PET, Δ is the slope of the saturation-to-vapour pressure curve (Pa K-1), γ is the psychrometric constant (0.066 kPa C-1) and α is the Priestley-Taylor coefficient (1.26). We derived Rn from the ERA5-Land product and G (soil heat flux) was set as zero at monthly time steps (Fisher et al., 2008). Following Li et al. (2013), we applied the nonlinear least squares regression method to derive the best-fit w for the study region and then applied the calibrated w to the estimation of ET_{Fu}.

2.4. Water balance evaluation

In this study, ET_{EB} estimates from multiple sources as well as an ensemble of them were compared with ET_{WB} in 53 catchments in central-western Europe. The climate in all catchments is relatively humid with a mean AI of 0.90, compared with the definition of arid catchments with an AI > 2 (Koppa and Gebremichael, 2017). To distinguish the impact of climate on water balances, we further differentiated eight catchments located in water-limited Budyko space (AI ≥ 1) from the remaining 45 energy-limited catchments: see Fig. 4a. The comparison of ET_{EB}-ET_{WB} in eight water-limited catchments can provide additional insight for the divergence of energy- and water-balance ET in energy-limited catchments. Given all components in the water balance calculation are independent, the overall uncertainty for ET_{WB} can be calculated as:

$$\sigma_{ET} = \sqrt{\sigma_P{}^2 + \sigma_Q{}^2 + \sigma_{TWSA}{}^2} \tag{4}$$

We used the correlation coefficient (r), relative bias (rBias), root mean square error (RMSE), and the Kling-Gupta efficiency scores (KGE) as our evaluation metrics. The KGE ranging from -∞ to 1 is a comprehensive indicator for model performance in hydrology, which integrates correlation, bias, and relative variability into a single performance statistic (Gupta et al., 2009). A higher KGE value indicates the simulations are closer to reproducing observations (Knoben et al., 2019). The evaluation metrics were computed as follows:

$$r = \frac{\sum_{i=1}^{N}(A_i - \overline{A})(B_i - \overline{B})}{\sqrt{\sum_{i=1}^{N}(A_i - \overline{A})^2 \sum_{i=1}^{N}(B_i - \overline{B})^2}} \tag{5}$$

$$rBias = \frac{\sum_{i=1}^{N}(A_i - B_i)}{\sum_{i=1}^{N} A_i} * 100\% \tag{6}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(A_i - B_i)^2}{N}} \tag{7}$$

$$KGE = 1 - \sqrt{(r-1)^2 + \left(\frac{\sigma_A}{\sigma_B} - 1\right)^2 + \left(\frac{\mu_A}{\mu_B} - 1\right)^2} \tag{8}$$

where N represents the sample size; A is the ET_{EB} estimate; and B is the reference ET that comprises ET_{WB}, ET_{Fu}, and upscaled ET; σ is the standard deviation of sample; μ is the mean value the sample. The subscript i represents different catchments in long-term assessment, whereas i represents the sample number of yearly and monthly data in annual and monthly assessment.

3. Results

3.1. Assessing spatial patterns of long-term average ET

Fig. 3 shows the statistical results of the comparison between long-term average ET_{WB} and ET_{EB} estimates from the four data sources along with the ensemble mean of these ET_{EB} products. Generally, the long-term average TWSA for all 53 catchments ranges from −10 to 15 mm yr^{-1} (see Fig. 6b) and is much lower than the long-term average ET of approximately 500 to 700 mm yr^{-1}. Consequently, TWSA is usually assumed to be negligible on a long-term basis. With respect to estimating long-term average ET in catchments that span a limited range of climate regimes, all ET_{EB} methods showed weak correlations ($r < 0.48$) with ET_{WB}; the correlation coefficient between ET_{WB} and the ET_{EB} ensemble was 0.45. As a comparison, when regarding TWSA as negligible, the correlation coefficient between ET_{PQ} and the ET_{EB} ensemble decreased to 0.40, which indicates that the inclusion of TWSA slightly improves the consistency between energy- and water-balance ET at the mean annual scale. Overall, large divergences were observed between the spatial variability of ET_{WB} and ET_{EB} in all selected catchments and taking GRACE-based TWSA into consideration can only marginally improve water budgets closure at the mean annual scale.

To explore the impact of different climates on the consistency between energy- and water-balance ET, we further conducted a comparison of ET_{WB} and ET_{EB} in energy-limited and water-limited catchments (Fig. 3). In energy-limited catchments, there was a weak correlation ($r = 0.37$) between ET_{WB} and the ET_{EB} ensemble, but in water-limited catchments, the correspondence between ET_{WB} and the ET_{EB} ensemble was significantly better ($r = 0.65$, $p < 0.05$). The agreement in terms of KGE was also higher in water-limited catchments (KGE = 0.56), compared with the analysis of ET_{WB} and the ET_{EB} ensemble in energy-limited catchments (KGE = 0.37). More importantly, the impact of

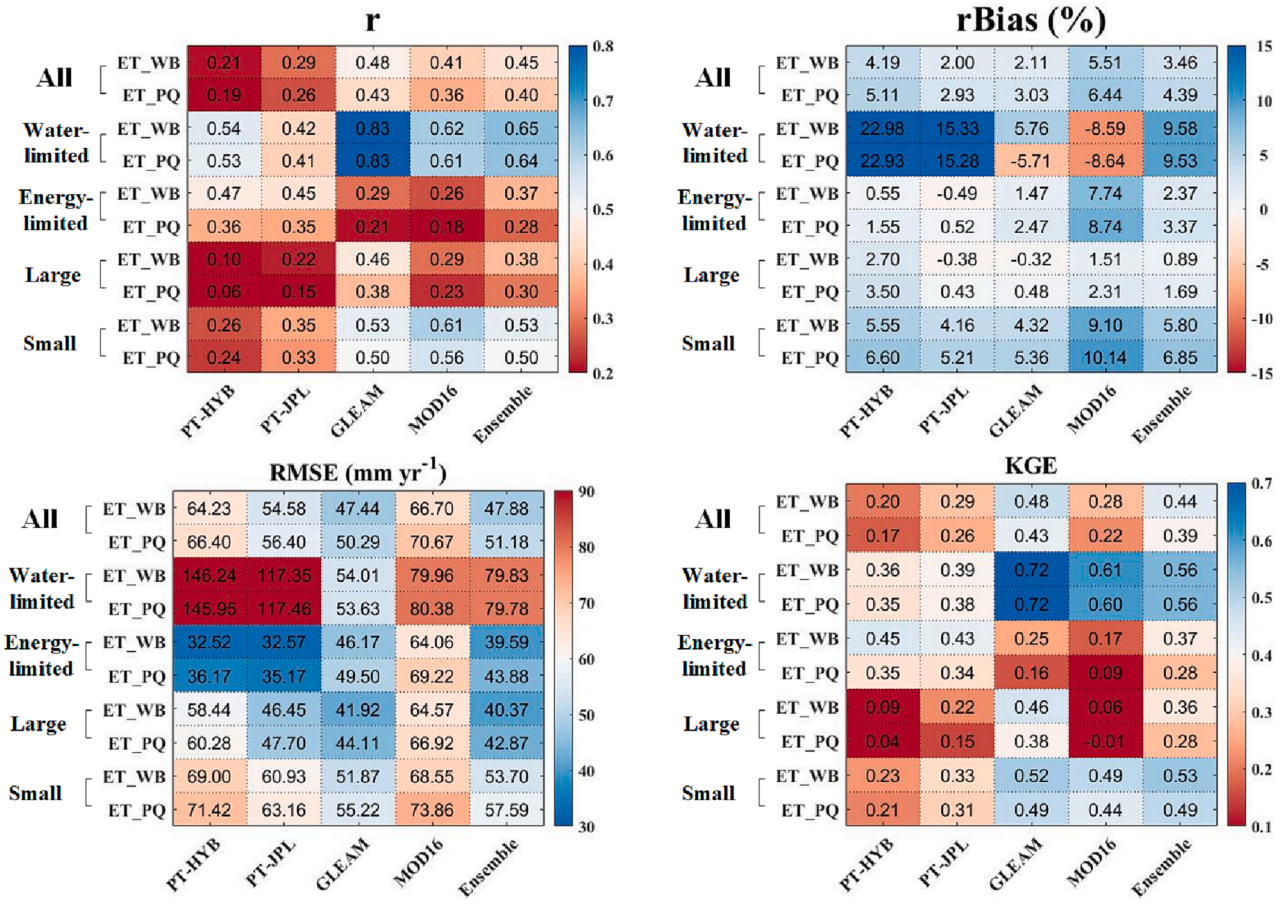

Fig. 3. Evaluation metrics for the comparison between ET_{WB} (as well as ET_{PQ}) and ET_{EB} for different groups: all 53 catchments; eight water-limited catchments, 45 energy-limited catchments; 25 large catchments; 28 small catchments.

incorporating GRACE-based TWSA into the water balance calculation on closing terrestrial water budgets is noticeable in energy-limited catchments, with a ΔKGE of +0.09. Among individual ET_{EB} products, MOD16 ET had the lowest agreement with the spatial patterns of ET_{WB} (KGE = 0.17) in energy-limited catchments, followed by GLEAM (KGE = 0.25). As a comparison, the spatial distribution of ET from PT-based methods had slightly improved correspondence with ET_{WB} in energy-limited catchments, with KGEs higher than 0.43. In water-limited catchments, most ET_{EB} methods (except for PT-JPL) reasonably captured the spatial changes of ET_{WB} in water-limited catchments and GLEAM yielded the best correlation (r = 0.83). Moreover, considering only 25 large-area catchments did not effectively improve the agreement between ET_{WB} and the ET_{EB} ensemble at the mean annual scale, with a lower divergence (ΔRMSE = −13 mm yr^{-1}) but a decreased correlation (Δr = −15%) than in the 28 small-area catchments. In conclusion, energy-limited catchments exhibited poor agreement in the spatial patterns of ET_{WB} and ET_{EB}, while significantly better correlations were found in water-limited catchments.

To understand the weak correlation between ET_{EB} and ET_{WB} in energy-limited catchments, we also calculated the water-balance inferred ET from the Budyko framework, which takes both water and energy constraints into consideration and has proven to efficiently estimate long-term ET from precipitation partitioning. Fig. 4a presents the derivation of the best-fit Budyko curve ($w = 3.61$) for all selected catchments using the nonlinear least squares regression method. We then applied the calibrated w to Fu's equation and compared ET_{Fu} and ET_{EB} in all selected catchments (Fig. 4b). Unlike ET_{WB}, long-term average ET_{Fu} estimates coincided well with ET_{EB} from multiple sources, exhibiting the highest correlation (r = 0.87) with the ET_{EB} ensemble across all 53 catchments. In energy-limited catchments, ET_{Fu} estimates had significantly better agreement with ET_{EB} from multiple sources (r > 0.86, KGE > 0.64), compared to the poor agreement between ET_{WB} and ET_{EB} in Fig. 3. Furthermore, both energy-limited catchments (r = 0.94) and water-limited catchments (r = 0.85) showed good correlation values between ET_{WB} and ET_{EB}. By contrast, the correlation coefficients between ET_{WB} and the ET_{EB} ensemble exhibited greater differences in energy-limited catchments (r = 0.36) and water-limited catchments (r = 0.65). Additionally, in water-limited catchments, there was a relatively poor correlation (r = 0.55) between PT-JPL ET and ET_{Fu}, and the MOD16 method yielded the largest discrepancy with ET_{Fu} (RMSE = 195 mm yr^{-1}) (Fig. 5e).

Since ET_{EB} estimates had good agreement with ET_{Fu} but diverged significantly from ET_{WB}, we further examined the relationship between the imbalance bias (Δ) and uncertainties associated with water cycle components to understand possible reasons behind the divergence of energy- and water-balance ET (Fig. 5). Our results illustrated that on a mean annual basis, precipitation as the largest component of water balances had the largest uncertainty across all catchments, with an average error of 81 mm yr^{-1}. By contrast, GRACE-based TWSA and *in situ* discharge had errors of 24 mm yr^{-1} and 12 mm yr^{-1}, respectively. Given that the catchment ID was ordered based on the drainage area, Fig. 5 demonstrates that the catchment area did not effectively change the uncertainties associated with GRACE-based TWSA and closure errors in water budgets at the mean annual scale, which was consistent with the results in Fig. 3. Moreover, through a comparison between ET_{WB} uncertainty and the imbalance bias, we found there were 18 catchments with imbalance bias exceeding the ET_{WB} uncertainty. This suggests that achieving the closure of water budgets is challenging for some

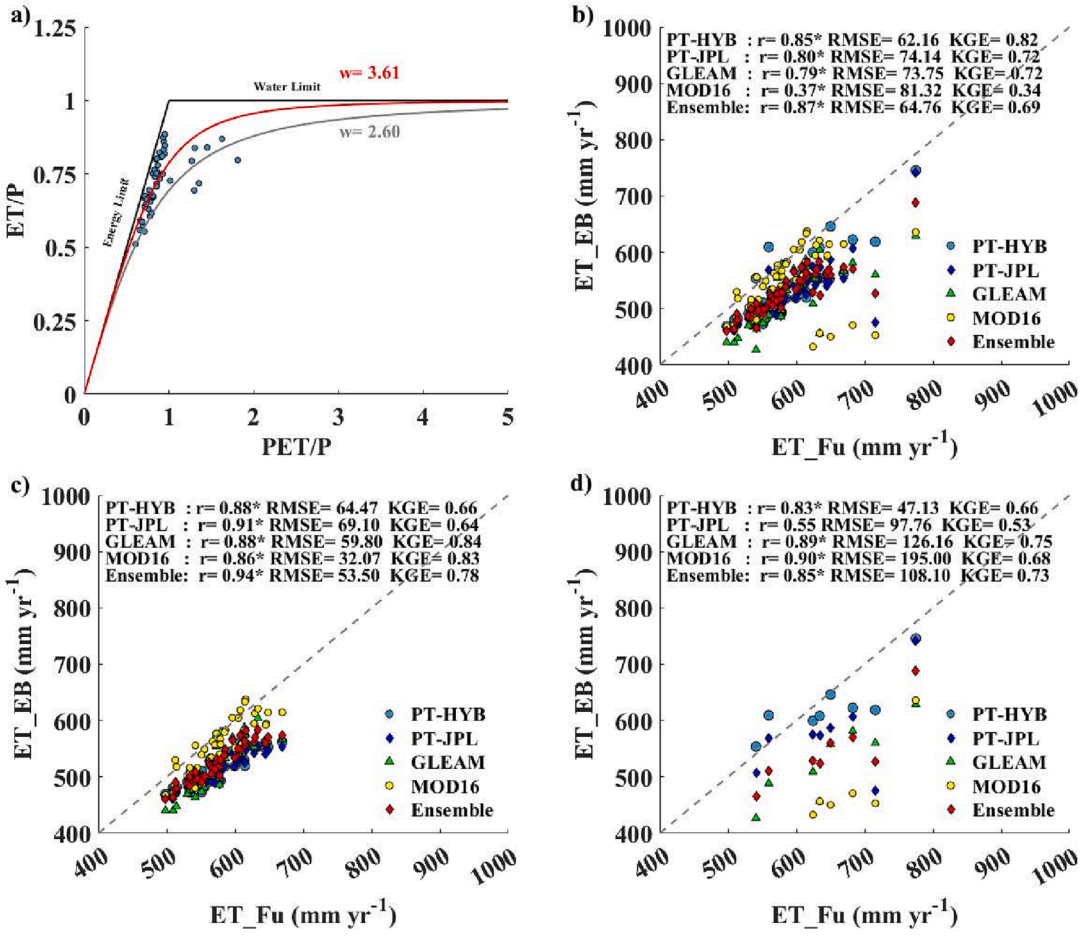

Fig. 4. The Budyko curve analysis using long-term averaged annual evaporative index (ET/P) and the Aridity Index with each point corresponding to a catchment (a), and simulated mean annual ET_{Fu} from the Budyko framework versus ET_{EB} for all 53 catchments (b) for the 45 energy-limited catchments (c), and for the 8 water-limited catchments (d). Asterisks indicate significant correlations ($p < 0.05$).

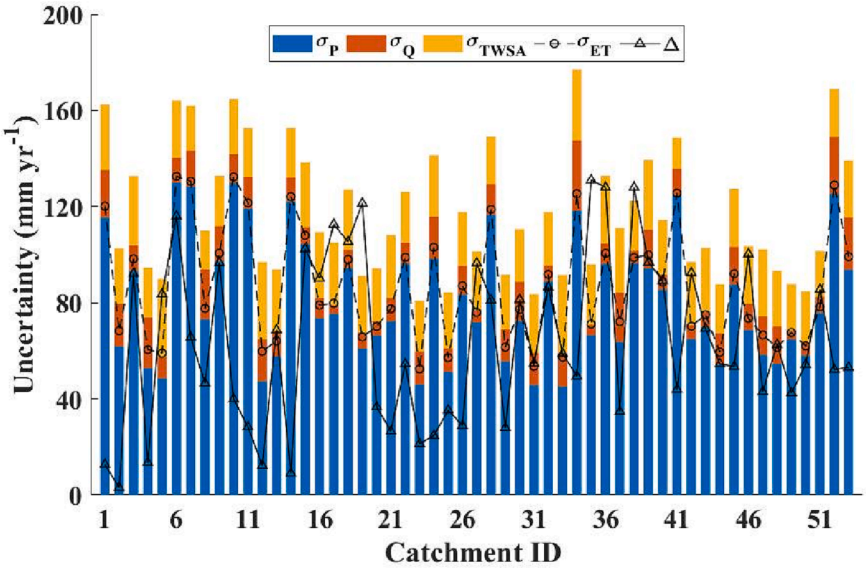

Fig. 5. The error budgets and the imbalance bias calculated from the ET_{EB} ensemble for all 53 catchments (ordered left to right from smallest to largest catchment).

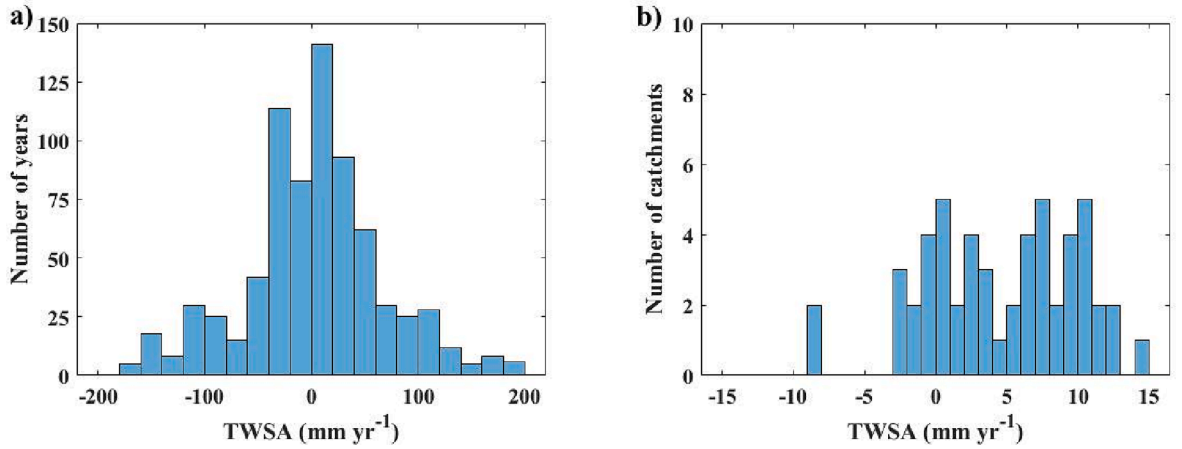

Fig. 6. Histogram of annual TWSA for all 53 catchments (a) and histogram of long-term TWSA for all 53 catchments (b).

catchments, even when using the ET_{EB} ensemble and considering random errors in water balance inputs.

3.2. Assessing interannual ET variability

To ascertain whether TWSA can be neglected in annual water balances, we first plotted the histogram of annual TWSA for all catchments (Fig. 6a). The analysis showed that annual TWSA mostly ranged between −200 and 200 mm yr^{-1}, which was considerably higher than the range of long-term TWSA averages (Fig. 6b). Consequently, TWSA is expected to play an important role in annual water balances. Then, we analysed the difference in the interannual variability of energy- and water-balance ET for energy-limited and water-limited catchments in Fig. 7. In energy-limited catchments, considering TWSA effectively improved the consistency between energy- and water-balance ET estimates at an annual timestep, and the correlation for the ET_{EB} ensemble increased from 0.07 (ET_{PQ}) to 0.37 (ET_{WB}). Conversely, in water-limited catchments, including TWSA into the water balance calculation did not enhance the ability of ET_{EB} methods to reproduce ET_{WB} time series: the correlation for the ET_{EB} ensemble decreased from 0.31 (ET_{PQ}) to 0.20 (ET_{WB}). More importantly, when using upscaled ET to reproduce ET_{WB} time series, the upscaled ET had almost no correlation with ET_{PQ} but moderate correlation with ET_{WB} ($r = 0.38$) across all 53 catchments, which highlights the uncertainty within ET_{WB} on an annual basis. The poor performance of ET_{EB} as well as upscaled ET to capture the inter-annual variability of ET_{WB} time series can be partly attributed to the lack of statistically significant correlation coefficients in many catchments. For instance, significant correlation coefficients ($p < 0.05$) between ET_{WB} and ET_{EB} were obtained in 13 catchments for GLEAM, five catchments for MOD16, four catchments for PT-JPL and three catchments for PT-hybrid. Consequently, these statistical results of correlation analyses should be interpreted with caution. In conclusion, it is commonly challenging for ET_{EB} products to capture the inter-annual variation of ET_{WB} and incorporating GRACE-based TWSA can partially improve the consistency between ET_{WB} and ET_{EB} in energy-limited catchments on an annual basis.

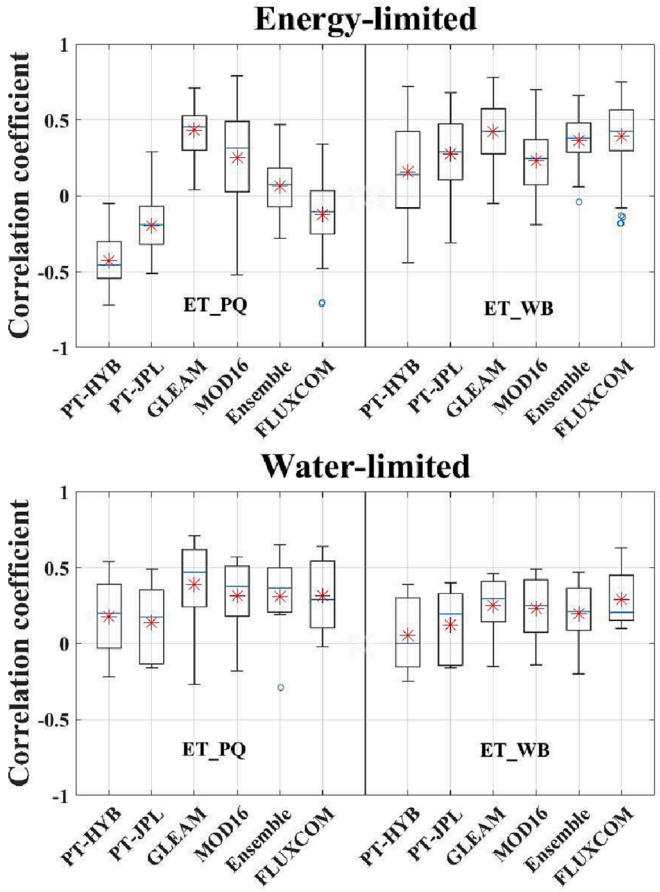

Fig. 7. Boxplot for the correlation coefficient between the interannual variation of ET_{WB} and ET_{EB} for the 45 energy-limited catchments and the 8 water-limited catchments. Inside the boxplot, the left part is the correlation analysis with ET_{PQ}, and the right part is the correlation analysis with ET_{WB}. Boxplots depict median, 25% to 75% range (box) and 10% to 90% range (whiskers). Asterisks indicate the means.

Fig. 8 displays the inter-annual variations of ET_{EB} and ET_{WB} for two energy-limited catchments and two water-limited catchments. Because annual ET varies greatly in magnitude across catchments, for better comparability we depict not the actual value of ET but the annual time series as ET anomalies. As shown in Fig. 8, both ET_{EB} estimates from multiple sources and upscaled ET from FLUXCOM failed to capture the inter-annual variability of ET_{WB}, which exhibited much greater variance. The substantial differences in the magnitude of ET variances highlight the difficulty of using ET_{EB} estimates to close the water budgets at the annual scale. Furthermore, irrespective of the magnitude of the interannual variability, the inter-annual variations in ET_{WB} were also not consistent with that of ET_{EB} in many basins. For instance, in the Przedborz catchment, MOD16 showed a high positive anomaly in 2010, whereas in the same year ET_{WB} displayed a strong negative anomaly. Similarly, for the Versen catchment in 2004, ET_{WB} was positive, whereas ET_{EB} estimates from all data sources were negative. Overall, all above results demonstrate the inability of ET_{EB} as well as upscaled ET to

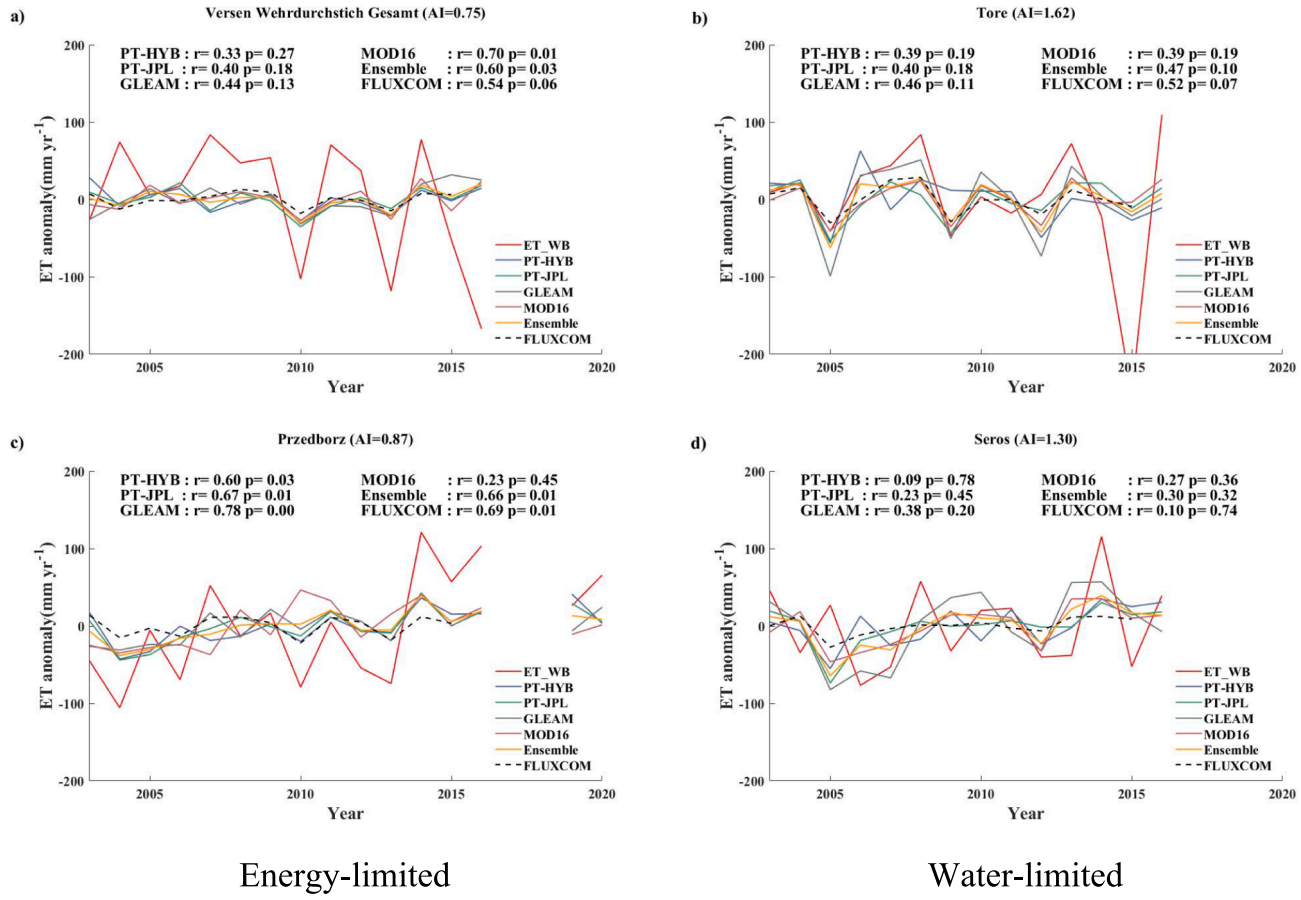

Fig. 8. Interannual variations of ET_{WB} anomaly and ET_{EB} anomaly in two energy-limited catchments (Versen Wehrdurchstich Gesamt and Przedberz) and two water-limited catchments (Tore and Seros). Due to the limited time span of the FLUXCOM data, the correlation statistics is calculated for the time period of 2003–2015.

capture the inter-annual variability of ET_{WB}, in terms of both magnitude and consistency, and the annual time series of ET_{WB} are characterized as markedly high variances.

To investigate the potential factors contributing to the extremely high variances of ET_{WB} and large discrepancy between annual ET_{WB} and ET_{EB}, we firstly calculated the closure errors in water budgets (Δ = P-Q-ET_{EB}) at an annual timestep. Instead of directly considering precipitation, we introduced the concept of excess precipitation (P-PET), which represents cumulative precipitation exceeding energy demand (Williams et al., 2012). After the reduction of PET, any increment in additional precipitation exceeding energy demand would not lead to increased ET but could introduce extra errors in ET_{WB}. Then, we analysed the relationship between the variations in closure errors and the dynamics of excess precipitation and GRACE-based TWSA. Fig. 9 demonstrates that the annual variations in closure errors retrieved from the ET_{EB} ensemble coincided well with excess precipitation in all 53 catchments with a mean r of 0.71, which revealed the close relationship between the closure errors and the highly variable precipitation. GRACE-based TWSA also correlated well with the closure errors from the ET ensemble across all 53 catchments (mean r = 0.52), but had mean bias of 100 mm yr^{-1} against closure errors in water budgets, indicating that incorporating GRACE-based TWSA did not thoroughly solve the closure errors at the annual scale.

3.3. Assessing ET at a monthly timestep

To evaluate how ET estimates compared at the monthly scale, Fig. 10 exhibited analyses of ET_{WB} and ET_{EB} for energy-limited and water-limited catchments. In energy-limited catchments, ET_{EB} estimates from all data sources had good correlations with ET_{WB} at the monthly scale, with r values ranging from 0.72 to 0.78. Among individual ET_{EB} methods, ET estimates from GLEAM yielded the highest correlation of 0.78 and the lowest RMSE of 22.68 mm $month^{-1}$, followed by MOD16 (r = 0.74, RMSE = 25.55 mm $month^{-1}$). In terms of KGE, the agreement between ET_{WB} and ET_{EB} was slightly lower for PT-hybrid (KGE = 0.65) and PT-JPL (KGE = 0.62), compared to GLEAM (KGE = 0.71). As opposed to the findings of the analyses conducted at long-term and annual scales, monthly ET_{EB} and ET_{WB} were less correlated in water-limited catchments than in energy-limited catchments. In water-limited catchments, ET_{EB} from GLEAM had the best agreement with monthly ET_{WB} (KGE = 0.55), followed by PT-hybrid (KGE = 0.42). However, ET_{EB} estimates from PT-hybrid were significantly higher than ET_{WB} with a rBias of 16.05% and MOD16 also exhibited large divergence from ET_{WB} with rBias of −16.79%.

To investigate how closely ET_{EB} agree with ET_{WB}, we plotted the monthly time series of ET_{WB} and ET_{EB} estimates from multiple sources for two energy-limited catchments and two water-limited catchments (Fig. 11). In general, ET_{EB} estimates from all sources reasonably captured the monthly variations of ET_{WB}, but they exhibited varying degrees of sensitivities to rainfall, which had sharp rises and falls. For example, extremely high ET_{WB} in summer months corresponded to heavy rainfall events, such as the rapid increment of ET_{WB} in 2011 for the Przedborz catchment (Fig. 11c) and in 2008 for the Seros catchment (Fig. 11d). However, ET_{EB} estimates mostly failed to capture the abrupt increment in ET_{WB} and exhibited an obvious underestimation of the ET_{WB} peak in summer months. Albeit ET_{EB} and ET_{WB} had coherently similar seasonality, the intra-annual variability of ET_{EB} was less pronounced than that of ET_{WB}, which closely followed the intra-annual variability of precipitation.

The improved correlation between ET_{EB} and ET_{WB} on the monthly

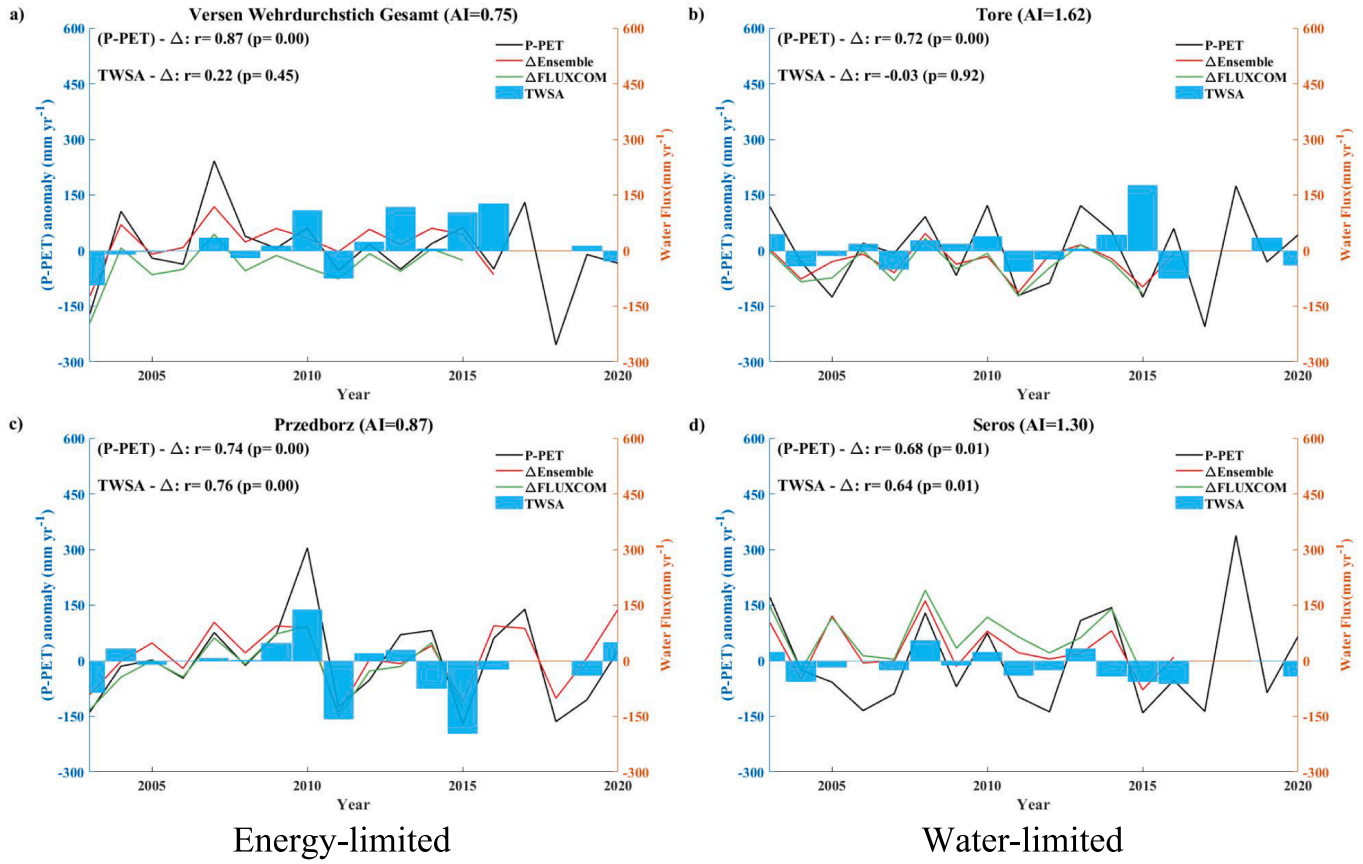

Fig. 9. Interannual variations of the P anomaly (left y axis) and the interannual variations of the closure errors of water budgets and TWSA (right y axis) in four representative catchments. For the correlation analysis, the Δ is calculated from the ET_{EB} ensemble.

scale may likely be caused by the consistent and coherent phase of seasonal cycles, compared with long-term and annual scales. To evaluate this, Fig. 12 shows the timing in the seasonal cycle of ET_{WB} and ET_{EB} ensemble. Statistical analysis demonstrates a strong correlation (r = 0.94) for the seasonality of ET_{WB} and the ET_{EB} ensemble across all 53 catchments and relatively better agreement was obtained in catchments with stronger precipitation seasonality, e.g. Tore (Fig. 12b) and Przedborz catchments (Fig. 12c). We further analysed the relationship between excess precipitation and the closure errors in water budgets at the seasonal scale to assess the impact of extremely heavy rainfall on sub-annual water balances. As shown in Fig. 12, good correlations between excess precipitation and closure errors were obtained in 26 catchments (r = 0.73), e.g. Versen and Seros catchments. By contrast, catchments with a marked precipitation seasonality e.g. Tore (Fig. 12b) and Przedborz catchments (Fig. 12c), exhibited higher correlations between ET_{WB} and ET_{EB}, which indicates that seasonality appears to buffer the impact of excess precipitation on water balances. Overall, despite uncertainties afflicting GRACE data in small-scale catchments, the disparities between ET_{WB} and ET_{EB} were found to closely related to excess precipitation and the similarity in the seasonal patterns seems to improve the consistency of energy- and water-balance ET at a monthly timestep.

4. Discussion

4.1. The large discrepancy between ET_{EB} and ET_{WB}

Despite the advances in observation techniques and satellite sensors, closing surface water budgets using multiple ET_{EB} products remains challenging. Our study revealed significant discrepancies between ET_{WB} and ET_{EB} from multiple sources on the mean annual scale, when TWSA is small enough to be omitted. Although incorporating GRACE-based TWSA into the water balance calculation slightly improves water budget closure, substantial divergences still exist in annual estimates of energy- and water-balance ET. These findings align with previous water budget studies that were unable to close terrestrial water budgets using analysed ET_{EB} products (Lehmann et al., 2022; Lorenz et al., 2014). Contrary to long-term and annual time scales, a better agreement was obtained between ET_{WB} and ET_{EB} on a monthly basis, but ET_{EB} exhibited less pronounced intra-annual variability compared to ET_{WB}, particularly in summer months with heavy rainfall (Springer et al., 2014). Overall, the consistent and significant discrepancies between energy- and water-balance ET estimates across multiple timesteps emphasize the need to investigate possible causes for the imbalance bias of water budgets.

The closure errors of water budgets can be explicitly attributed to the uncertainties associated with ET_{WB}, which is computed as the residual of water balance equation. The primary source of errors in ET_{WB} stem from precipitation data quality. As precipitation is the largest component in terrestrial water budgets, uncertainties at the same relative level would contribute to larger absolute errors than *in situ* runoff and GRACE-based TWSA (Xu et al., 2022). Although the density of precipitation gauges in central-western Europe facilitates bias correction for precipitation datasets, precipitation obtained from different sources still exhibit large divergence at catchment scale. Fig. 5 confirms that precipitation errors consistently account for the largest part in ET_{WB} errors across 53 catchments, compared with TWSA and runoff. In an attempt to minimize errors and get a robust catchment precipitation, we used the ensemble mean of precipitation from gauge-based E-OBS dataset, ERA5-Land reanalysis dataset and WorldClim v2.1 dataset. Additional sources of precipitation data could be tested in future to further reduce errors, even if their availability is currently limited to national-scale products. These include products derived from terrestrial radar measurements or

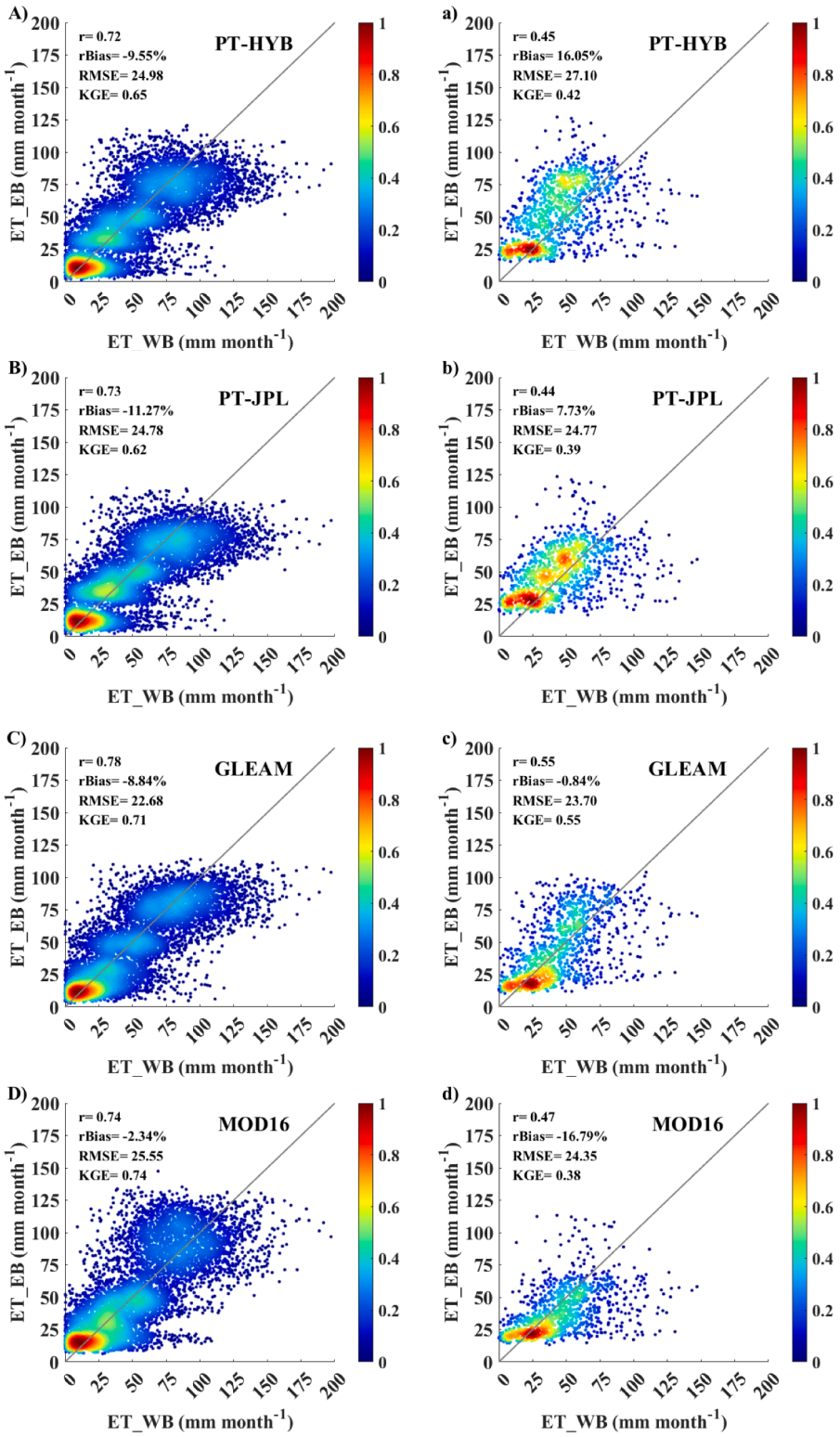

Fig. 10. Comparison of the ET_{WB} to ET_{EB} at the monthly scale for energy-limited catchments (left) and for water-limited catchments (right). The gray line denotes 1:1 line.

microwave links from cellular network provider, which have proven to be an additional source of rainfall information to complement traditional rainfall data (Graf et al., 2020; Sun et al., 2018).

A second source of error in ET_{WB} can be attributed to the coarse-resolution GRACE data, which has a large spatial mismatch with small-scale catchments (Wiese et al., 2016). Previous studies attempted to incorporate water storage predictions obtained from land surface models or microwave soil moisture into the spatial downscaling of GRACE data, but these two storage proxies are limited by human induced change in water storage (e.g. irrigation and groundwater extraction) and storage dynamics occurring in deeper layers of the unsaturated zone, respectively (Crow et al., 2017; Pascolini-Campbell

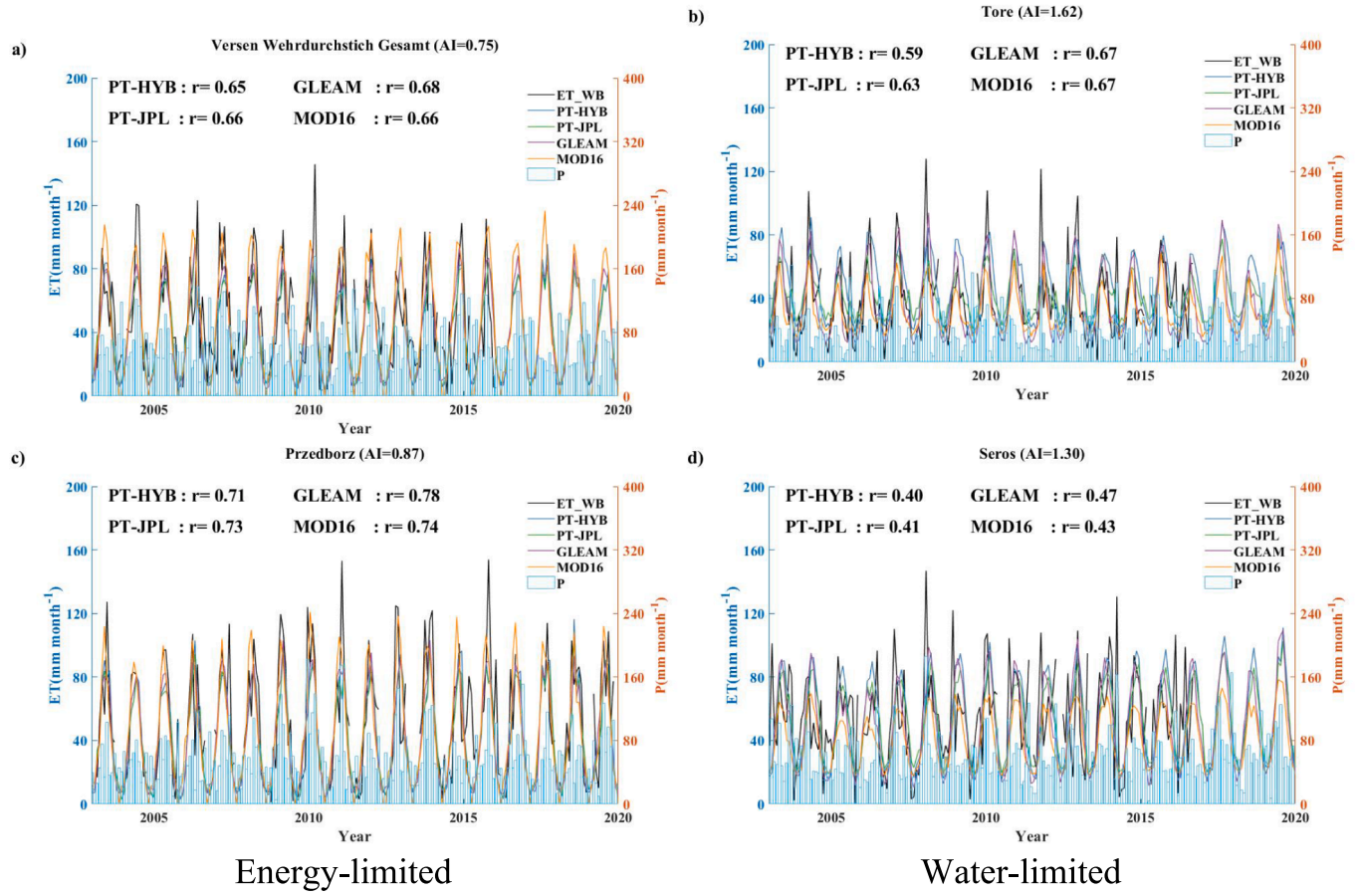

Fig. 11. Time series of monthly ET_{WB}, ET_{EB} from multiple data sources (left y axis), and precipitation dynamics in two energy-limited catchments (Versen Wehrdurchstich Gesamt and Przedberz) and two water-limited catchments (Tore and Seros).

et al., 2021). In our study, we used the average of three GRACE products resampled to the common 25 km resolution, but the limited drainage area will introduce leakage and attenuation uncertainties in modelled TWSA (Scanlon et al., 2016). We note that the catchment area, varying from 1,000 to 150,000 km^2, did not significantly affect the closure errors in water budgets. This is consistent with the work of Zhang et al. (2012), who found similar correspondences between ET_{WB} and ET_{EB} in two groups of catchments: 107 small-area catchments (<5,000 km^2) and 90 large-area catchments (>5,000 km^2). In our opinion, although the accuracy of TWSA is inevitably limited by GRACE data in small-scale catchments, the consistent and considerable discrepancies between energy- and water-balance ET indicate that TWSA is not the key reason for the differences in the long-term spatial pattern of ET and the different amplitudes of annual ET time series. That is because terrestrial water storage is more stable than other water balance components and consequently TWSA is usually assumed to be negligible at longer time-steps. However, for shorter temporal scales (e.g., monthly) terrestrial water storage becomes more variable and the uncertainty of TWSA plays a more important role in ET_{WB} (Zhong et al., 2020).

Another factor contributing to the imbalance bias of water budgets is the intricate nature of groundwater baseflow. The assumption of negligible groundwater flow at long-term scale holds for continental drainage basins, in which surface drainage coincides with groundwater flow divides (Rodell et al., 2004). However, in smaller catchments with a humid climate, lateral groundwater loss/gain can be significant, especially when heavy rainfall events and anthropogenic withdrawals occur (Le Moine et al., 2007). In our study, we applied a threshold of runoff-rainfall coefficients to exclude runoff-dominant catchments, but when non-negligible lateral groundwater flow occurs, the groundwater baseflow crossing topographic boundaries cannot be adequately resolved by TWSA. Hence, the difficulty of accurately quantifying groundwater baseflow is another plausible explanation for the unclosed water balances at finer spatial scales.

Uncertainties in ET_{EB} estimates also contribute to closure errors in water budgets. Although ET_{EB} methods are expected to have good performance in humid regions, their accuracy are still limited due to the sensitivity to chosen algorithm and the quality of input data (Ershadi et al., 2015; Yao et al., 2019). With respect to the model structure, the GLEAM product, contrary to the other products, explicitly considers the influence of soil water stress on ET through microwave-based measurements within a simple water balance framework (Bai and Liu, 2018). Consequently, the GLEAM product outperforms other ET_{EB} sources in closing terrestrial water budgets, particularly for water-limited catchments. In contrast, ET_{EB} methods such as MOD16 and PT-JPL use surface air humidity as a proxy for soil wetness, which proves problematic under dry condition (Gao et al., 2016; Yao et al., 2013). The PT-hybrid model, unlike the original PT-JPL method, uses a compound SWIR-microwave index to parameterise fsm and has a better correlation ($\Delta r = 12\%$) with mean annual ET_{WB} in water-limited catchments (Zhang et al., 2021). With respect to the forcing data quality, previous studies have found that the chosen net radiation data account for the largest part of differences in ET_{EB} (Anderson et al., 2019; Badgley et al., 2015). Systematic biases in reanalysis net radiation, influenced by clouds and aerosols, may explain the poorer performance of PT-based methods than the GLEAM product in reproducing annual ET_{WB} time series. Despite efforts to resolve errors in individual ET_{EB} method through an ensemble mean of multi-source ET_{EB}, significant divergences still exist between ET_{WB} and the ET_{EB} ensemble across time scales.

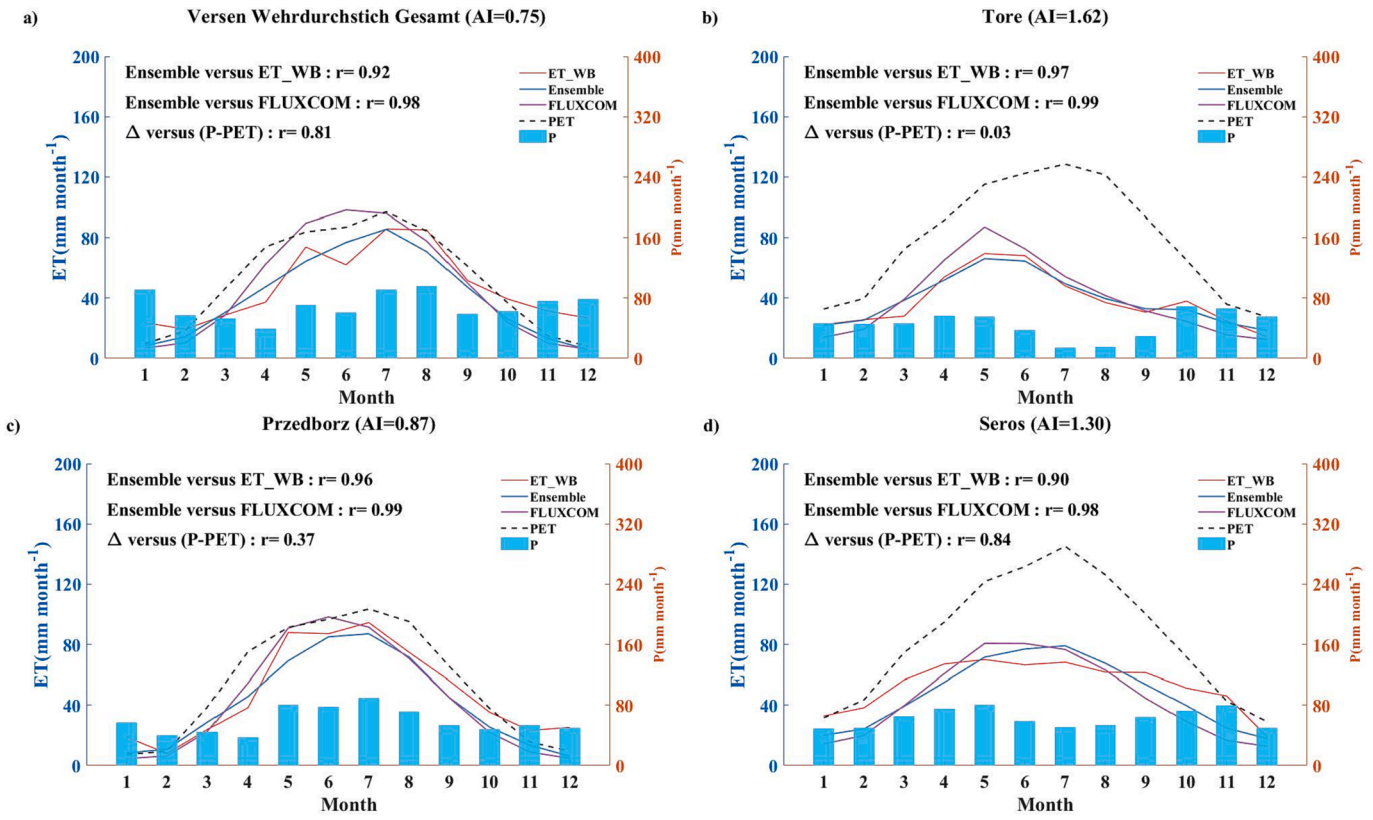

Fig. 12. Seasonal cycles of ET_{WB}, the ET_{EB} ensemble, upscaled ET from FLUXCOM, PET (left y axis), and precipitation (right y axis) in four representative catchments.

4.2. Factors affecting the ET_{WB}–ET_{EB} consistency across multiple timesteps

Although ET_{EB} methods consistently exhibit large discrepancies in closing terrestrial water budgets, our analysis reveals varying degrees of consistency between ET_{EB} and ET_{WB} across multiple timesteps, particularly for energy-limited catchments. On a monthly basis, ET_{EB} estimates were in close correspondence with ET_{WB}: correlations ranged from 0.72 to 0.78 in energy-limited catchments and from 0.44 to 0.55 in water-limited catchments. These statistical results are in line with those of Ruhoff et al. (2022), who reported similar correlation values against ET_{WB}, which ranged from 0.45 to 0.60 for eight global ET products across fifty catchments. However, on a mean annual basis, ET_{EB} from multiple sources cannot capture the spatial patterns of ET_{WB} with an average r of 0.35. Moreover, substantially lower consistency were observed in energy-limited catchments compared with water-limited catchments at the mean annual scale, which is consistent with the finding of Carter et al. (2018) that ET_{EB} exhibits higher long-term divergence from ET_{WB} in humid regions with decreased covariability of ET_{EB} and ET_{WB}. Generally, several factors may contribute to the inconsistency between ET_{EB} and ET_{WB}; below, we discuss possible factors affecting ET_{EB}-ET_{WB} correlations across multiple timesteps.

On a mean annual basis, the poor correlation between energy- and water-balance ET estimates in energy-limited catchments can be attributed to uncertainties associated with ET_{WB}, mainly arising from precipitation errors. Our results revealed that ET_{EB} models poorly captured the spatial changes in long-term ET_{WB} in energy-limited catchments (mean r = 0.37) but yielded good correlations in water-limited catchments (mean r = 0.60). Similar results were reported by Pang et al. (2021) who found that long-term distributions of ET_{WB} and ET_{EB} exhibited an correlation of 0.55 in 22 semi-arid catchments. The different outcomes of ET_{EB}-ET_{WB} analyses can be explained by the fact that under different climatic conditions, the primary factor controlling ET is different: available water for water-limited regions; available energy for energy-limited regions (Liu et al., 2015; Yang et al., 2006). On the one hand, precipitation plays a dominant role in determining the amount of available water and further regulates the spatial variability of ET in water-limited regions (Zhang et al., 2004). Consequently, good correlations were obtained for the long-term spatial patterns of ET_{WB} and ET_{EB} in water-limited catchments. On the other hand, ET_{WB} calculated as the residual of terrestrial water balance is directly influenced by precipitation errors and diverges significantly from ET_{EB} in energy-limited catchments (Soni and Syed, 2021). Water supply is mostly abundant in energy-limited regions and the primary factor regulating ET is available energy rather than precipitation. Given that TWSA is negligible at long-term scale, precipitation errors accrue in ET_{WB} and contributes to the poor correlation between energy- and water-balance ET estimates. In contrast, water-balance inferred ET_{Fu} from the Budyko framework, which incorporates both water and energy constraints into ET estimation, exhibits significantly better agreement with ET_{EB}. Overall, the substantial differences between energy- and water-balance ET estimates can be attributed to the uncertainty of ET_{WB} in energy-limited catchments. In such catchments, considering energy constraints for ET, like the Budyko framework, or resolving ET_{WB} errors by using more accurate precipitation datasets has the potential to narrow the discrepancy between mean annual ET_{WB} and ET_{EB} for energy-limited catchments.

For annual assessment, the inability of ET_{EB} to capture the interannual variations of ET_{WB} can be predominantly attributed to errors in ET_{WB}, which was found to closely follow the temporal variability of precipitation. Our results illustrated that the interannual variability of ET_{WB} correlated weakly with the ET_{EB} ensemble in all 53 catchments (r = 0.34). Even for large-scale catchments with drainage area larger than GRACE footprint, it has been observed that there is a lack of correlation in annual time series of ET_{WB} and ET_{EB} (Bai and Liu, 2018; Liu et al., 2016). With respect to the high variances of ET_{WB}, Pang et al. (2021)

identified precipitation as the primary driver for the interannual variability of ET_{WB}, accounting for 81.0% of the trend in semi-arid regions, followed by net radiation (42.7%) and wind speed (25.3%). Ukkola and Prentice (2013) also found that precipitation accounts for 54–55% of the interannual variations in ET_{WB} for wet catchments and 94–95% in dry catchments. When precipitation is either too high or too low, ET_{WB} closely follows the fluctuations in precipitation and diverges more from ET_{EB}. As a result, inconsistencies were observed for the interannual variability of ET_{WB} and ET_{EB} in most catchments, which is consistent with Carter et al. (2018). Further analysis in Fig. 9 revealed the close relationship between the closure errors in water budgets and additional precipitation exceeding energy demand (mean r = 0.71 for all 53 catchments). To our knowledge, precipitation beyond energy demand cannot contribute to the increment in ET and the transform of excess precipitation to runoff and TWSA has large uncertainties, which makes ET_{WB} problematic to be compared with ET_{EB}. Given that errors from modelled TWSA cannot totally explain the high variances of ET_{WB}, the highly variable precipitation may culprit in the significant divergence between ET_{EB} and ET_{WB} at the annual scale.

The intra-annual variations of energy- and water-balance ET exhibited similar seasonal patterns and yielded stronger correlations than at long-term and annual scales, particularly for energy-limited catchments. This improved correlation at the monthly scale might be attributable to the similar seasonal patterns, which are influenced by a combination of climatic characteristics, such as precipitation, solar radiation, air temperature, humidity and wind speed (Christoffersen et al., 2014). The dominant factor affecting the hydrological cycle varies greatly across temporal scales, with seasonality, particularly of precipitation, playing a more important role in affecting water balances at shorter time scales compared to long-term water balances controlled by the aridity index (Fu and Wang, 2019; Wang and Tang, 2014). The combined influence of climatic seasonality leads to the consistency in seasonal cycles of ET, contributing to good agreement between ET_{WB} and ET_{EB} at a monthly timestep. Moreover, we also found that ET_{EB} had closer agreement with ET_{WB} in catchments with stronger precipitation seasonality (Ruhoff et al., 2022). Conversely, in catchments with weak precipitation seasonality, the closure errors in water budgets closely correlated with excess precipitation and relatively weak consistency exists between ET_{WB} and ET_{EB}. Our finding confirms the work of Sahoo et al. (2011), who found that the monthly imbalance bias showed a seasonal cycle and can be mostly attributed to precipitation uncertainty. In water-limited catchments, we found that ET_{WB} and ET_{EB} diverged more, which is the exact opposite of our findings in the long-term average analysis. This result may be explained by the lower accuracy of ET_{EB} models in water-limited regions with short vegetation canopies and the increased importance of human activities (e.g. irrigation) at shorter timesteps (Ershadi et al., 2014). Overall, the agreement in the seasonal patterns may contribute to the better consistency between energy- and water-balance ET estimates at the monthly scale compared to long-term and annual scales.

4.3. Limitations of our water balance evaluation

There are other sources of uncertainty in our water balance evaluation. Firstly, errors may arise from the scale mismatch between ET_{EB} products and basin-wide ET_{WB} estimates. To facilitate comparison, we uniformly interpolated all gridded datasets to a common 5 km spatial resolution except for GARCE data. However, this spatial interpolation process inevitably introduces uncertainties in the comparison between ET_{WB} and ET_{EB} across temporal scales (Miralles et al., 2016). Secondly, in regards to the poor correlation between energy- and water-balance ET estimates at the mean annual scale, the limited range of multiyear ET (500–700 mm yr^{-1}) in all 53 catchments is not favorable to have a good correlation performance. Therefore, we utilised multiple metrics, e.g., KGE, to evaluate the agreement between ET_{EB} and ET_{WB}. Incorporating more catchments that span a wide range of climate regimes could make the correlation analysis of ET_{EB}-ET_{WB} more robust, because it is relatively 'easy' to capture the large ET variation across catchments of widely varying characteristics (Zhu et al., 2022). Thirdly, including eight water-limited catchments aimed to provide inference information for the divergence of energy- and water-balance ET estimates in energy-limited catchments, whereas the limited number of samples (eight catchments with a mean AI of 1.39) would underrepresent water-limited catchments. Moreover, as for the comparison at annual scale, the limited length ($\leq$18 years) and data gaps in certain datasets, such as FLUXCOM (unavailable after 2015) and GRACE data (11 missing months in 2017–2018), result in non-significant correlations between annual time series of ET_{WB} and those of ET_{EB} in many catchments.

To provide insights of the interpretation of our water balance evaluation, we incorporated upscaled ET from FLUXCOM and ET_{Fu} from the Budyko framework to find possible explanations for the divergence of energy- and water-balance ET. On the one hand, in comparison to the poor correlations between ET_{WB} and ET_{EB} at long-term scale, the good correspondence between ET_{EB} and ET_{Fu} suggests that uncertainty associated with ET_{WB} is also responsible for the significant divergences between ET_{WB} and ET_{EB}. Upscaled ET also fails to capture the interannual variations of ET_{WB}, likely due to the highly variable precipitation. The above findings suggest that the non-closure issue of water budgets, particularly in energy-limited catchments, limits the application of ET_{WB} to benchmark ET_{EB}. Similarly, Han et al. (2015) attributed the decline they found in ET_{EB}-ET_{WB} correlations to the poor accuracy of ET_{WB} in wet climates. Based on the inconsistency between ET_{WB} and atmospheric-inferred ET_{WB}, Li et al. (2019) suggested using atmospheric-inferred ET_{WB} instead as benchmark in runoff-dominant catchments. Taken together, we argue that errors from precipitation and TWSA data introduce large uncertainties in ET_{WB} in humid regions, contributing to poor agreement with ET_{EB} estimates at long-term and annual scales. On the other hand, although upscaled ET from FLUXCOM has been commonly used as a benchmark in carbon and water studies, the high accuracy of machine learning-based upscaling methods relies on having sufficient *in situ* ET observations as inputs (Jung et al., 2019; Miralles et al., 2016). In data-sparse regions, caution is needed when applying upscaled ET as reference due to the limited representativeness of flux towers. Moreover, machine learning models lack physical mechanisms and provide limited insights into the evaporation process (Gentine et al., 2018). Combining machine learning methods with semi-empirical or physical methods will not only provide useful information for the validation of ET estimates but also contributes to the enhancement of ET models from remote sensing. With respect to the robust Budyko framework, it tends to perform best for large-scale catchments and longer time-frames (Li et al., 2022). To be applied at finer spatial scales, the spatial variability between catchments should be considered in the Budyko framework, due to the impacts of soil properties, vegetation and topography on water balances (Bai et al., 2020). To be applied at shorter timesteps, the non-steady state of water storage should be considered in the Budyko framework and short-term climate variability as well as extreme events are expected to affect water balances (Fu and Wang, 2019). In our study, we calibrated the best-fit w for the study region using historical records of precipitation and discharge and then applied the Fu's equation to predict long-term ET. Given that the inclusion of energy constraints seems to improve the closure of terrestrial water budgets, further work is needed to investigate if the modified Budyko framework with increased model complexity or reconstructed TWSA that takes the impact of temperature on water balances into consideration has the potential to improve bias-correction of ET_{WB} at shorter timesteps (Bai et al., 2022).

5. Conclusions

This study comprehensively evaluated the divergence of energy- and water-balance ET estimates across 53 catchments in central-western Europe. ET_{Fu} retrievals from the Budyko framework and upscaled ET

from FLUXCOM product were incorporated to the interpretation of water balance evaluation. The results revealed significant disparities between ET_{WB} and ET_{EB} estimates from multiple sources especially for energy-limited catchments, regardless of the inclusion of GRACE-based TWSA. Despite the consistent and considerable closure errors in water budgets, the consistency between ET_{EB} and ET_{WB} varies significantly with time scales. At the long-term scale, ET_{EB} diverged significantly from the spatial patterns of both ET_{PQ} calculated as precipitation minus runoff and ET_{WB} considering GRACE-based TWSA, whereas improved agreement was achieved between ET_{EB} and water-balance inferred ET_{Fu}. On an annual basis, both ET_{EB} and upscaled ET could not capture the interannual variability of ET_{WB} with non-significant correlations in most catchments. Meanwhile, the closure errors of water budgets are highly variable in time and closely follow excess precipitation beyond energy demand. Unlike the long-term and annual time scale, monthly ET_{EB} correlated well with ET_{WB} and the intra-annual variability of ET_{WB} and ET_{EB} followed similar seasonal cycles. Although excess precipitation in heavy-rainfall months still influence the agreement between ET_{WB} and ET_{EB}, the similarity in seasonal patterns contribute to good performance in correlation analysis, particular for catchments with a marked precipitation seasonality.

Our analyses shed light on the uncertainties inherent in the water-balance approach. They emphasize that using ET_{WB} as a benchmark without accounting for uncertainties within the water balance calculation can result in biased conclusions about the performance of ET_{EB} in humid regions. Although errors from coarse-resolution GRACE data are likely larger at finer spatial scales, the inclusion of GRACE-based TWSA partially improves the consistency of ET_{WB} and ET_{EB} in small-scale catchments, presumably because terrestrial water storage is more spatially homogenous than other water cycle components. On a mean annual basis, TWSA is small enough to be omitted and substantial divergences still exist between energy- and water-balance ET, indicating the GRACE-based TWSA is not the primary cause for the differences in long-term spatial patterns of ET_{EB} and ET_{PQ}. Precipitation is highly variable in time and space and predominantly regulates the spatio-temporal variability of ET_{WB} in humid regions. However, for such regions, precipitation variability does not always directly affect ET and the quantification of the transition from excess precipitation to runoff or TWSA has large uncertainties. Consequently, the closure errors of water budgets were found to closely related to additional precipitation beyond energy demand. In summary, because of errors in precipitation and TWSA data, care must be taken when conducting water balance evaluation across temporal scales, especially for humid regions where available energy is the primary factor limiting ET. In such regions, the robustness of ET_{WB} needs to be improved by using enhanced-quality input data or - like the Budyko framework - taking energy constraints for the evaporation process into consideration.

CRediT authorship contribution statement

Lilin Zhang: Conceptualization, Methodology, Data curation, Writing – original draft, Writing – review & editing, Visualization, Funding acquisition. **Michael Marshall:** Conceptualization, Validation, Writing – review & editing, Supervision. **Anton Vrieling:** Conceptualization, Formal analysis, Writing – review & editing, Supervision. **Andrew Nelson:** Conceptualization, Software, Writing – review & editing, Project administration.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This study was supported by the China Scholarship Council (CSC) under Grant 201806040218. Thanks to Dr. Joy Burrough-Boenisch, who was the language editor of a near-final draft of the paper. The authors also would like to express their gratitude to the researchers and their teams for providing all of the data used in this study and the last gratitude to Xiangyi Bei for the assistance in writing the paper.

References

Albergel, C., et al., 2018. ERA-5 and ERA-Interim driven ISBA land surface model simulations: which one performs better? Hydrol. Earth Syst. Sci. 22 (6), 3515–3532.
Anderson, M., et al., 2019. Impact of insolation data source on remote sensing retrievals of evapotranspiration over the California Delta. Remote Sens. (Basel) 11 (3), 216.
Anderson, M.C., Norman, J.M., Mecikalski, J.R., Otkin, J.A., Kustas, W.P., 2007. A climatological study of evapotranspiration and moisture stress across the continental United States based on thermal remote sensing: 1. Model formulation. J. Geophys. Res. Atmos. 112 (D10).
Badgley, G., Fisher, J.B., Jiménez, C., Tu, K.P., Vinukollu, R., 2015. On uncertainty in global terrestrial evapotranspiration estimates from choice of input forcing datasets. J. Hydrometeorol. 16 (4), 1449–1455. https://doi.org/10.1175/jhm-d-14-0040.1.
Bai, H., et al., 2022. Evaluation of evapotranspiration for exorheic basins in China using an improved estimate of terrestrial water storage change. J. Hydrol. 610, 127885 https://doi.org/10.1016/j.jhydrol.2022.127885.
Bai, P., Liu, X., 2018. Intercomparison and evaluation of three global high-resolution evapotranspiration products across China. J. Hydrol. 566, 743–755. https://doi.org/10.1016/j.jhydrol.2018.09.065.
Bai, P., Liu, X., Zhang, D., Liu, C., 2020. Estimation of the Budyko model parameter for small basins in China. Hydrol. Process. 34 (1), 125–138. https://doi.org/10.1002/hyp.13577.
Biggs, T., Petropoulos, G., Velpuri, N., Marshall, M., 2015. Remote sensing of actual evapotranspiration from croplands, Remote Sensing Handbook: Remote Sensing of Water Resources, Disasters, and Urban Studies. CRC Press, pp. 59–99.
Boergens, E., Dobslaw, H., Dill, R., 2019. GFZ GravIS RL06 continental water storage anomalies.
Bowen, I.S., 1926. The ratio of heat losses by conduction and by evaporation from any water surface. Phys. Rev. 27 (6), 779–787. https://doi.org/10.1103/PhysRev.27.779.
Brust, C., et al., 2021. Using SMAP Level-4 soil moisture to constrain MOD16 evapotranspiration over the contiguous USA. Remote Sens. Environ. 255, 112277 https://doi.org/10.1016/j.rse.2020.112277.
Budyko, M.I., 1974. Climate and Life. Academic press.
Carter, E., Hain, C., Anderson, M., Steinschneider, S., 2018. A water balance-based, spatiotemporal evaluation of terrestrial evapotranspiration products across the contiguous United States. J. Hydrometeorol. 19 (5), 891–905. https://doi.org/10.1175/jhm-d-17-0186.1.
Castle, S.L., et al., 2016. Remote detection of water management impacts on evapotranspiration in the Colorado River Basin. Geophys. Res. Lett. 43 (10), 5089–5097. https://doi.org/10.1002/2016GL068675.
Christoffersen, B.O., et al., 2014. Mechanisms of water supply and vegetation demand govern the seasonality and magnitude of evapotranspiration in Amazonia and Cerrado. Agric. For. Meteorol. 191, 33–50. https://doi.org/10.1016/j.agrformet.2014.02.008.
Cornes, R.C., van der Schrier, G., van den Besselaar, E.J.M., Jones, P.D., 2018. An ensemble version of the E-OBS temperature and precipitation data sets. J. Geophys. Res. Atmos. 123 (17), 9391–9409. https://doi.org/10.1029/2017JD028200.
Crow, W.T., Han, E., Ryu, D., Hain, C.R., Anderson, M.C., 2017. Estimating annual water storage variations in medium-scale (2000–10 000 km2) basins using microwave-based soil moisture retrievals. Hydrol. Earth Syst. Sci. 21 (3), 1849–1862. https://doi.org/10.5194/hess-21-1849-2017.
D'Odorico, P., et al., 2018. The global food-energy-water Nexus. Rev. Geophys. 56 (3), 456–531. https://doi.org/10.1029/2017RG000591.
Ershadi, A., McCabe, M., Evans, J.P., Chaney, N.W., Wood, E.F., 2014. Multi-site evaluation of terrestrial evaporation models using FLUXNET data. Agric. For. Meteorol. 187, 46–61.
Ershadi, A., McCabe, M.F., Evans, J.P., Wood, E.F., 2015. Impact of model structure and parameterization on Penman-Monteith type evaporation models. J. Hydrol. 525, 521–535. https://doi.org/10.1016/j.jhydrol.2015.04.008.
Fick, S.E., Hijmans, R.J., 2017. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. Int. J. Climatol. 37 (12), 4302–4315. https://doi.org/10.1002/joc.5086.
Fisher, J.B. et al., 2020. ECOSTRESS: NASA's next generation mission to measure evapotranspiration from the International Space Station. Water Resour. Res., 56(4): e2019WR026058.
Fisher, J.B., Tu, K.P., Baldocchi, D.D., 2008. Global estimates of the land–atmosphere water flux based on monthly AVHRR and ISLSCP-II data, validated at 16 FLUXNET sites. Remote Sens. Environ. 112 (3), 901–919.

Fu, B., 1981. On the calculation of the evaporation from land surface. Sci. Atmos. Sin 5 (1), 23–31.
Fu, J., Wang, W., 2019. On the lower bound of Budyko curve: The influence of precipitation seasonality. J. Hydrol. 570, 292–303.
Gao, Y., Gan, G., Liu, M., Wang, J., 2016. Evaluating soil evaporation parameterizations at near-instantaneous scales using surface dryness indices. J. Hydrol. 541, 1199–1211. https://doi.org/10.1016/j.jhydrol.2016.08.027.
Gentine, P., Pritchard, M., Rasp, S., Reinaudi, G., Yacalis, G., 2018. Could machine learning break the convection parameterization deadlock? Geophys. Res. Lett. 45 (11), 5742–5751. https://doi.org/10.1029/2018GL078202.
Ghilain, N., Arboleda, A., Gellens-Meulenberghs, F., 2011. Evapotranspiration modelling at large scale using near-real time MSG SEVIRI derived data. Hydrol. Earth Syst. Sci. 15 (3), 771–786. https://doi.org/10.5194/hess-15-771-2011.
Glenn, E.P., Huete, A.R., Nagler, P.L., Hirschboeck, K.K., Brown, P., 2007. Integrating remote sensing and ground methods to estimate evapotranspiration. Crit. Rev. Plant Sci. 26 (3), 139–168. https://doi.org/10.1080/07352680701402503.
Graf, M., Chwala, C., Polz, J., Kunstmann, H., 2020. Rainfall estimation from a German-wide commercial microwave link network: optimized processing and validation for 1 year of data. Hydrol. Earth Syst. Sci. 24 (6), 2931–2950. https://doi.org/10.5194/hess-24-2931-2020.
Grdc, 2011. Watershed Boundaries of GRDC Stations/Global Runoff Data Centre. Federal Institute of Hydrology (BfG), Koblenz, Germany.
Gupta, H.V., Kling, H., Yilmaz, K.K., Martinez, G.F., 2009. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. J. Hydrol. 377 (1), 80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003.
Han, E., Crow, W.T., Hain, C.R., Anderson, M.C., 2015. On the use of a water balance to evaluate interannual terrestrial ET variability. J. Hydrometeorol. 16 (3), 1102–1108. https://doi.org/10.1175/jhm-d-14-0175.1.
Han, J. et al., 2020. Assessing the steady-state assumption in water balance calculation across global catchments. Water Resour. Res., 56(7): e2020WR027392. 10.1029/2020WR027392.
He, Y., Wang, K., Feng, F., 2021. Improvement of ERA5 over ERA-Interim in Simulating Surface Incident Solar Radiation throughout China. J. Clim. 34 (10), 3853–3867. https://doi.org/10.1175/jcli-d-20-0300.1.
Hobbins, M.T., Ramírez, J.A., Brown, T.C., 2001. The complementary relationship in estimation of regional evapotranspiration: An enhanced advection-aridity model. Water Resour. Res. 37 (5), 1389–1403. https://doi.org/10.1029/2000WR900359.
Holmes, J.W., 1984. Measuring evapotranspiration by hydrological methods. Agric Water Manag 8 (1), 29–40. https://doi.org/10.1016/0378-3774(84)90044-1.
Hu, G., Jia, L., Menenti, M., 2015. Comparison of MOD16 and LSA-SAF MSG evapotranspiration products over Europe for 2011. Remote Sens. Environ. 156, 510–526. https://doi.org/10.1016/j.rse.2014.10.017.
Jiang, C., Ryu, Y., 2016. Multi-scale evaluation of global gross primary productivity and evapotranspiration products derived from Breathing Earth System Simulator (BESS). Remote Sens. Environ. 186, 528–547. https://doi.org/10.1016/j.rse.2016.08.030.
Jung, M., et al., 2010. Recent decline in the global land evapotranspiration trend due to limited moisture supply. Nature 467 (7318), 951.
Jung, M., et al., 2011. Global patterns of land-atmosphere fluxes of carbon dioxide, latent heat, and sensible heat derived from eddy covariance, satellite, and meteorological observations. J. Geophys. Res. Biogeo. 116 (G3).
Jung, M., et al., 2019. The FLUXCOM ensemble of global land-atmosphere energy fluxes. Sci. Data 6 (1), 74. https://doi.org/10.1038/s41597-019-0076-8.
Jung, M., Reichstein, M., Bondeau, A., 2009. Towards global empirical upscaling of FLUXNET eddy covariance observations: validation of a model tree ensemble approach using a biosphere model. Biogeosciences 6 (10), 2001–2013. https://doi.org/10.5194/bg-6-2001-2009.
Kim, D., Choi, M., Chun, J.A., 2022. Linking the complementary evaporation relationship with the Budyko framework for ungauged areas in Australia. Hydrol. Earth Syst. Sci. 26 (23), 5955–5969. https://doi.org/10.5194/hess-26-5955-2022.
Knoben, W.J.M., Freer, J.E., Woods, R.A., 2019. Technical note: Inherent benchmark or not? Comparing Nash-Sutcliffe and Kling-Gupta efficiency scores. Hydrol. Earth Syst. Sci. 23 (10), 4323–4331. https://doi.org/10.5194/hess-23-4323-2019.
Koppa, A., Gebremichael, M., 2017. A framework for validation of remotely sensed precipitation and evapotranspiration based on the Budyko hypothesis. Water Resour. Res. 53 (10), 8487–8499. https://doi.org/10.1002/2017WR020593.
Kyaw Tha Paw, U., Qiu, J., Su, H.-B., Watanabe, T., Brunet, Y., 1995. Surface renewal analysis: a new method to obtain scalar fluxes. Agric. For. Meteorol. 74 (1), 119–137. https://doi.org/10.1016/0168-1923(94)02182-J.
Le Moine, N., Andréassian, V., Perrin, C., Michel, C., 2007. How can rainfall-runoff models handle intercatchment groundwater flows? Theoretical study based on 1040 French catchments. Water Resour. Res. 43 (6) https://doi.org/10.1029/2006WR005608.
Lehmann, F., Vishwakarma, B.D., Bamber, J., 2022. How well are we able to close the water budget at the global scale? Hydrol. Earth Syst. Sci. 26 (1), 35–54. https://doi.org/10.5194/hess-26-35-2022.
Li, X., et al., 2019. Evapotranspiration estimation for Tibetan Plateau headwaters using conjoint terrestrial and atmospheric water balances and multisource remote sensing. Water Resour. Res. 55 (11), 8608–8630. https://doi.org/10.1029/2019WR025196.
Li, S., Du, T., Gippel, C.J., 2022. A modified Fu (1981) equation with a time-varying parameter that improves estimates of inter-annual variability in catchment water balance. Water Resour. Manag. 36 (5), 1645–1659. https://doi.org/10.1007/s11269-021-03057-1.
Li, D., Pan, M., Cong, Z., Zhang, L., Wood, E., 2013. Vegetation control on water and energy balance within the Budyko framework. Water Resour. Res. 49 (2), 969–976. https://doi.org/10.1002/wrcr.20107.
Liu, Y., et al., 2015. Evapotranspiration in Northern Eurasia: Impact of forcing uncertainties on terrestrial ecosystem model estimates. J. Geophys. Res. Atmos. 120 (7), 2647–2660. https://doi.org/10.1002/2014JD022531.
Liu, W., et al., 2016. A worldwide evaluation of basin-scale evapotranspiration estimates against the water balance method. J. Hydrol. 538, 82–95. https://doi.org/10.1016/j.jhydrol.2016.04.006.
Lorenz, C., et al., 2014. Large-scale runoff from landmasses: A global assessment of the closure of the hydrological and atmospheric water balances. J. Hydrometeorol. 15 (6), 2111–2139. https://doi.org/10.1175/jhm-d-13-0157.1.
Ma, N., Szilagyi, J., Zhang, Y., 2021. Calibration-free complementary relationship estimates terrestrial evapotranspiration globally. Water Resour. Res., 57(9): e2021WR029691. 10.1029/2021WR029691.
Marshall, M., Funk, C., Michaelsen, J., 2012. Examining evapotranspiration trends in Africa. Clim. Dyn. 38 (9–10), 1849–1865.
Martens, B., et al., 2017. GLEAM v3: satellite-based land evaporation and root-zone soil moisture. Geosci. Model Dev. 10 (5), 1903–1925. https://doi.org/10.5194/gmd-10-1903-2017.
McCabe, M.F., Wood, E.F., 2006. Scale influences on the remote estimation of evapotranspiration using multiple satellite sensors. Remote Sens. Environ. 105 (4), 271–285. https://doi.org/10.1016/j.rse.2006.07.006.
Meijninger, W.M.L., et al., 2002. Determination of area-averaged water vapour fluxes with large aperture and radio wave scintillometers over a heterogeneous surface – Flevoland field experiment. Bound.-Lay. Meteorol. 105 (1), 63–83. https://doi.org/10.1023/A:1019683616097.
Michel, D., et al., 2016. TheWACMOS-ET project–Part 1: Tower-scale evaluation of four remote-sensing-based evapotranspiration algorithms. Hydrol. Earth Syst. Sci. 20 (2), 803–822.
Miralles, D.G., et al., 2011. Global land-surface evaporation estimated from satellite-based observations. Hydrol. Earth Syst. Sci. 15 (2), 453–469. https://doi.org/10.5194/hess-15-453-2011.
Miralles, D.G., et al., 2016. The WACMOS-ET project – Part 2: Evaluation of global terrestrial evaporation data sets. Hydrol. Earth Syst. Sci. 20 (2), 823–842. https://doi.org/10.5194/hess-20-823-2016.
Mu, Q., Heinsch, F.A., Zhao, M., Running, S.W., 2007. Development of a global evapotranspiration algorithm based on MODIS and global meteorology data. Remote Sens. Environ. 111 (4), 519–536.
Mu, Q., Zhao, M., Running, S.W., 2011. Improvements to a MODIS global terrestrial evapotranspiration algorithm. Remote Sens. Environ. 115 (8), 1781–1800.
Muñoz Sabater, J., 2019. ERA5-Land hourly data from 1981 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS).
Pan, Y., et al., 2017. Detection of human-induced evapotranspiration using GRACE satellite observations in the Haihe River basin of China. Geophys. Res. Lett. 44 (1), 190–199. https://doi.org/10.1002/2016GL071287.
Pan, S., et al., 2020. Evaluation of global terrestrial evapotranspiration using state-of-the-art approaches in remote sensing, machine learning and land surface modeling. Hydrol. Earth Syst. Sci. 24 (3), 1485–1509. https://doi.org/10.5194/hess-24-1485-2020.
Pang, X., et al., 2021. Long term variation of evapotranspiration and water balance based on upscaling eddy covariance observations over the temperate semi-arid grassland of China. Agric. For. Meteorol. 308–309, 108566 https://doi.org/10.1016/j.agrformet.2021.108566.
Pascolini-Campbell, M.A., Reager, J.T., Fisher, J.B., 2020. GRACE-based mass conservation as a validation target for basin-scale evapotranspiration in the contiguous United States. Water Resour. Res., 56(2): e2019WR026594. 10.1029/2019WR026594.
Pascolini-Campbell, M., Fisher, J.B., Reager, J.T., 2021. GRACE-FO and ECOSTRESS synergies constrain fine-scale impacts on the water balance. Geophys. Res. Lett., 48 (15): e2021GL093984. 10.1029/2021GL093984.
Pastorello, G., et al., 2020. The FLUXNET2015 dataset and the ONEFlux processing pipeline for eddy covariance data. Sci. Data 7 (1), 225. https://doi.org/10.1038/s41597-020-0534-3.
Rodell, M., et al., 2004. Basin scale estimates of evapotranspiration using GRACE and other observations. Geophys. Res. Lett. 31 (20) https://doi.org/10.1029/2004GL020873.
Rodell, M., et al., 2018. Emerging trends in global freshwater availability. Nature 557 (7707), 651–659. https://doi.org/10.1038/s41586-018-0123-1.
Ruhoff, A., et al., 2022. Global evapotranspiration datasets assessment using water balance in South America. Remote Sens. (Basel) 14 (11), 2526.
Sahoo, A.K., et al., 2011. Reconciling the global terrestrial water budget using satellite remote sensing. Remote Sens. Environ. 115 (8), 1850–1865. https://doi.org/10.1016/j.rse.2011.03.009.
Sakumura, C., Bettadpur, S., Bruinsma, S., 2014. Ensemble prediction and intercomparison analysis of GRACE time-variable gravity field models. Geophys. Res. Lett. 41 (5), 1389–1397. https://doi.org/10.1002/2013GL058632.
Sauer, V.B., Meyer, R., 1992. Determination of error in individual discharge measurements. 2331-1258, US Geological Survey; Books and Open-File Reports Section [distributor].
Save, H., Bettadpur, S., Tapley, B.D., 2016. High-resolution CSR GRACE RL05 mascons. J. Geophys. Res. Solid Earth 121 (10), 7547–7569. https://doi.org/10.1002/2016JB013007.
Scanlon, B.R., et al., 2016. Global evaluation of new GRACE mascon products for hydrologic applications. Water Resour. Res. 52 (12), 9412–9429. https://doi.org/10.1002/2016WR019494.
Senay, G.B., Friedrichs, M., Singh, R.K., Velpuri, N.M., 2016. Evaluating Landsat 8 evapotranspiration for water use mapping in the Colorado River Basin. Remote Sens. Environ. 185, 171–185. https://doi.org/10.1016/j.rse.2015.12.043.

Soni, A., Syed, T.H., 2021. Analysis of variations and controls of evapotranspiration over major Indian River Basins (1982–2014). Sci. Total Environ. 754, 141892 https://doi.org/10.1016/j.scitotenv.2020.141892.

Springer, A., Kusche, J., Hartung, K., Ohlwein, C., Longuevergne, L., 2014. New estimates of variations in water flux and storage over Europe based on regional (re) analyses and multisensor observations. J. Hydrometeorol. 15 (6), 2397–2417. https://doi.org/10.1175/jhm-d-14-0050.1.

Sun, Q., et al., 2018. A review of global precipitation data sets: data sources, estimation, and intercomparisons. Rev. Geophys. 56 (1), 79–107. https://doi.org/10.1002/2017RG000574.

Ukkola, A.M., Prentice, I.C., 2013. A worldwide analysis of trends in water-balance evapotranspiration. Hydrol. Earth Syst. Sci. 17 (10), 4177–4187. https://doi.org/10.5194/hess-17-4177-2013.

Velpuri, N.M., Senay, G.B., Singh, R.K., Bohms, S., Verdin, J.P., 2013. A comprehensive evaluation of two MODIS evapotranspiration products over the conterminous United States: Using point and gridded FLUXNET and water balance ET. Remote Sens. Environ. 139, 35–49. https://doi.org/10.1016/j.rse.2013.07.013.

Vinukollu, R.K., Meynadier, R., Sheffield, J., Wood, E.F., 2011. Multi-model, multi-sensor estimates of global evapotranspiration: climatology, uncertainties and trends. Hydrol. Process. 25 (26), 3993–4010. https://doi.org/10.1002/hyp.8393.

Wan, Z., et al., 2015. Water balance-based actual evapotranspiration reconstruction from ground and satellite observations over the conterminous United States. Water Resour. Res. 51 (8), 6485–6499. https://doi.org/10.1002/2015WR017311.

Wang, X., de Linage, C., Famiglietti, J., Zender, C.S., 2011. Gravity Recovery and Climate Experiment (GRACE) detection of water storage changes in the Three Gorges Reservoir of China and comparison with in situ measurements. Water Resour. Res. 47 (12) https://doi.org/10.1029/2011WR010534.

Wang, K., Dickinson, R.E., 2012. A review of global terrestrial evapotranspiration: Observation, modeling, climatology, and climatic variability. Rev. Geophys. 50 (2).

Wang, D., Tang, Y., 2014. A one-parameter Budyko model for water balance captures emergent behavior in darwinian hydrologic models. Geophys. Res. Lett. 41 (13), 4569–4577. https://doi.org/10.1002/2014GL060509.

Wang, C., Wang, S., Fu, B., Zhang, L., 2016. Advances in hydrological modelling with the Budyko framework: A review. Prog. Phys. Geogr. 40 (3), 409–430.

Wiese, D.N., Landerer, F.W., Watkins, M.M., 2016. Quantifying and reducing leakage errors in the JPL RL05M GRACE mascon solution. Water Resour. Res. 52 (9), 7490–7502. https://doi.org/10.1002/2016WR019344.

Williams, C.A., et al., 2012. Climate and vegetation controls on the surface water balance: Synthesis of evapotranspiration measured across a global network of flux towers. Water Resour. Res. 48 (6) https://doi.org/10.1029/2011WR011586.

Xu, J., Ma, Z., Yan, S., Peng, J., 2022. Do ERA5 and ERA5-land precipitation estimates outperform satellite-based precipitation products? A comprehensive comparison between state-of-the-art model-based and satellite-based precipitation products over mainland China. J. Hydrol. 605, 127353 https://doi.org/10.1016/j.jhydrol.2021.127353.

Yang, D., Sun, F., Liu, Z., Cong, Z., Lei, Z., 2006. Interpreting the complementary relationship in non-humid environments based on the Budyko and Penman hypotheses. Geophys. Res. Lett. 33 (18) https://doi.org/10.1029/2006GL027657.

Yao, Y., et al., 2013. MODIS-driven estimation of terrestrial latent heat flux in China based on a modified Priestley-Taylor algorithm. Agric. For. Meteorol. 171, 187–202.

Yao, Y., et al., 2017. Improving global terrestrial evapotranspiration estimation using support vector machine by integrating three process-based algorithms. Agric. For. Meteorol. 242, 55–74. https://doi.org/10.1016/j.agrformet.2017.04.011.

Yao, Y., et al., 2019. Evaluation of a satellite-derived model parameterized by three soil moisture constraints to estimate terrestrial latent heat flux in the Heihe River basin of Northwest China. Sci. Total Environ. 695, 133787 https://doi.org/10.1016/j.scitotenv.2019.133787.

Yin, L., Wang, X., Feng, X., Fu, B., Chen, Y., 2020. A comparison of SSEBop-model-based evapotranspiration with eight evapotranspiration products in the Yellow River Basin. China. Remote Sensing 12 (16), 2528.

Zhang, L., et al., 2004. A rational function approach for estimating mean annual evapotranspiration. Water Resour. Res. 40 (2) https://doi.org/10.1029/2003WR002710.

Zhang, Y., et al., 2010. Using long-term water balances to parameterize surface conductances and calculate evaporation at 0.05° spatial resolution. Water Resour. Res. 46 (5) https://doi.org/10.1029/2009WR008716.

Zhang, Y., et al., 2012. Decadal trends in evaporation from global energy and water balances. J. Hydrometeorol. 13 (1), 379–391. https://doi.org/10.1175/jhm-d-11-012.1.

Zhang, Y., et al., 2018. A Climate Data Record (CDR) for the global terrestrial water budget: 1984–2010. Hydrol. Earth Syst. Sci. 22 (1), 241–263. https://doi.org/10.5194/hess-22-241-2018.

Zhang, L., Marshall, M., Nelson, A., Vrieling, A., 2021. A global assessment of PT-JPL soil evaporation in agroecosystems with optical, thermal, and microwave satellite data. Agric. For. Meteorol. 306, 108455 https://doi.org/10.1016/j.agrformet.2021.108455.

Zhong, Y., Zhong, M., Mao, Y., Ji, B., 2020. Evaluation of evapotranspiration for exorheic catchments of China during the GRACE era: from a water balance perspective. Remote Sens. (Basel) 12 (3), 511.

Zhu, W., Tian, S., Wei, J., Jia, S., Song, Z., 2022. Multi-scale evaluation of global evapotranspiration products derived from remote sensing images: Accuracy and uncertainty. J. Hydrol. 611, 127982 https://doi.org/10.1016/j.jhydrol.2022.127982.