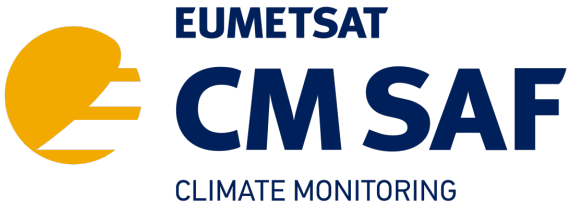
EUMETSAT Satellite Application Facility on Climate Monitoring



Use of CM SAF evapotranspiration product to assess the water balance in European basins.

Reference Number:	SAF/CM/DWD/CDOP4/REP/ET_WB_UC
Issue/Revision Index:	1.0
Date:	15.04.2025

Document Signature Table

	Name	Function	Signature	Date
Author	William Moutier	Author and CM SAF Scientist		15/04/25
Author	Miguel Barrios	Author and CM SAF Scientist		15/04/25
Author	Jan De Pue	Author and CM SAF Scientist		15/04/25
Author	Françoise Gellens-Meulenberghs	Author and CM SAF Scientist		15/04/25
Author	Alirio Arboleda	Author and CM SAF Scientist		15/04/25
Author	Nicolas Clerbaux	CM SAF International Board Member		15/04/25
Editor	Marc Schröder	CM SAF Science Coordinator		
Approval	CM SAF Steering Group			
Release	Rainer Hollmann	CM SAF Project Manager		

Distribution List

Internal Distribution	
Name	**No. Copies**
DWD / Archive	1

External Distribution		
Company	**Name**	**No. Copies**
Public		0

Document Change Record

Issue/ Revision	Date	DCN No.	Changed Pages/Paragraphs
1.0	01/02/2024		First version submitted.

Applicable Documents

Reference	Title	Code
AD 1		
AD 2		

Reference Documents

Reference	Title	Code
RD 1		
RD 2		

Table of Contents

List of Tables

List of Figures

Deleted: Table 1: Overview of the datasets used in this study, including their names, temporal and spatial resolution (as used in this study), and periods of availability. 14
Table 2: Summary information of the seven largest GRDC river stations (area (≥ 90 000 km^2) extracted from the dataset. The columns include the GRDC station number (grdc_no), river and station names, country code (ISO 3166), geographical coordinates (latitude and longitude in decimal degrees), catchment area (in km^2), and gauge altitude (in meters above sea level). 57

Deleted: Figure 1: Schematic representation of data sources and methodologies used for estimating and comparing evapotranspiration based on the water balance method including (ET_{WB}) or not (ET_{PO}) the change in water storage with remote sensing-based/reanalysis products. 13

Figure 2: Spatial distribution of the 342 river discharge stations selected from the Global Runoff Data Centre (GRDC) across Europe. Each dot represents a station, with red indicating large basins (area > 90 000 km^2) and black indicating small basins (area ≤ 90 000 km^2). The associated basin shapes are also shown, colored according to their aridity index (AI): blue for energy-limited basins (AI ≤ 1) and orange for water-limited basins (AI > 1). 15

Figure 3: Probability density functions (PDFs) of monthly evapotranspiration (ET) estimates for all selected basins, over the period 2004-2020, from the five regridded (0.25°) datasets: CM SAF (blue), LSA SAF (red), GLEAM (green), ERA5-Land (purple), and GLDAS-2.1 (yellow). Vertical solid lines indicate the median ET values for each dataset. 19

Figure 4: Probability density functions (PDFs) of monthly precipitation (P) estimates for all selected basins, over the period 2004-2020, from the three regridded (0.25°) datasets: GIRAFE (blue), MSWEP (red) and E-OBS (green). Vertical solid lines indicate the median precipitation values for each dataset. 21

Figure 5: Probability density functions (PDFs) of monthly water storage (dSdt ; mm month^{-1}) estimates for all selected basins, over the period 2004-2020, from the three sources: GRACE-CSR (blue), GRACE-JPL (red) and GRACE-GFZ (green). Vertical solid lines indicate the median values for each PDFs. 23

Figure 6: Probability density functions (PDFs) of the aridity index (AI), computed as the ratio of reference evapotranspiration from LSA SAF (ETrefLSA SAF) to precipitation (P), for all selected basins over the period 2004–2020. The analysis is based on three regridded (0.25°) precipitation datasets: GIRAFE (blue), E-OBS (green), and MSWEP (red). The grey-shaded area represents the distribution of median AI values derived from the three individual datasets. Vertical solid lines indicate the median AI for each dataset. 24

Figure 7: Location of the GRDC station Düsseldorf (red dot) and its associated river basin (blue polygon) derived from GRDC shapefiles (area of 147680 km^2). The basin, part of the Rhine River system, is classified as large (area > 90 000 km^2) and energy-limited (AI ≤ 1). 26

Figure 8: Monthly time-series of key hydrological components for the Rhine River at Düsseldorf (2003–2020). The panels display, from top to bottom: (1) observed discharge (Q_{GRDC}); (2) precipitation (P) from multiple datasets (GIRAFE, E-OBS, MSWEP), including their mean (P Average , red solid line); (3) water storage change (dSdt) estimated from GRACE solutions (CSR, JPL, GFZ), with their ensemble mean (red solid line); (4) evapotranspiration (ET) estimates from different datasets (CM SAF, LSA SAF, GLEAM, ERA5-Land, GLDAS 2.1), including their mean (red solid line); and (5) water balance-derived evapotranspiration with (ET_{WB}) and without (ET_{PO}) including the water storage component, along with the monthly averages of key hydrological variables used in their computation: discharge (Q_{GRDC}; red solid line), mean water storage ($dSdtGRACE_{Average}$; yellow solid line), mean precipitation ($P_{Average}$, bar-plot in blue) and the mean evapotranspiration from all independent datasets ($ET_{Average}$). 27

Figure 9: Scatter plots comparing monthly evapotranspiration estimates from the CM SAF product ($ET_{CM\ SAF}$ in mm month^{-1}) with water balance-derived evapotranspiration estimates with (ET_{WB} in mm month^{-1}; left panels) and without (ET_{PO} in mm month^{-1}; right panels) considering dSdt for all basins. GIRAFE (top panels), MSWEP (middle panels), and E-OBS (bottom panels) precipitation products are considered for the calculation of ET_{WB} and ET_{PO}. The dashed black line represents the 1:1 relationship. Statistical metrics, including the coefficient of determination (R^2), unbiased root mean square [1]

1 Introduction

In the context of global warming and rapid changes of the land use due to human economical activities, it is fundamental to be able to accurately estimate key variables such as the evapotranspiration (ET). Accurate quantification of ET is critical for understanding the terrestrial water cycle, carbon cycle, land–atmosphere interactions, surface energy balance, and for managing water resources, particularly under changing climatic conditions (Sellers et al., 1997; Oki and Kanae, 2006; Trenberth et al., 2009; Liou and Kar, 2014; Michel et al., 2016; Fisher et al., 2017; Behrendt et al., 2020). As pointed out in the IPCC report on “Climate Models and Their Evaluation” (Randall et al., 2007), ET is a key variable in model evaluation. The Coupled Model Intercomparison Project (CMIP, Eyring et al. (2016)), which evaluates the climate models and provides input on future climate predictions for the IPCC reports, relies on ET climate data records (CDRs). Numerous studies made a dedicated analysis of the modelled ET in the CMIP ensemble (Mueller and Seneviratne, 2014; Lian et al., 2018; Wang et al., 2021). Examples of impactful multi-model benchmarking and evaluation studies are legion, including the Protocol for the Analysis of Land Surface Models (PALS) Land Surface Model Benchmarking Evaluation Project (PLUMBER; Best et al. (2015)) and Global Energy and Water Cycle Exchanges (GEWEX) LandFlux project (McCabe et al., 2016). They highlight the importance of ET, and emphasise that the climate modelling community is strongly relying on ET CDRs. Accurate estimates of ET can also be useful for the calibration of hydrological models (Rajib et al., 2018; Nijzink et al., 2018; Sirisena et al., 2020). Moreover, various studies have used ET to investigate the impacts of climate change on hydrology (Liu et al., 2021), droughts (Joetzjer et al., 2013; Kim and Rhee, 2016; Cook et al., 2022; Perez et al., 2024; Zhang et al., 2023b), feedbacks with vegetation (Yang et al., 2023), and to assess planetary boundary layer (Wang-Erlandsson et al., 2022). In addition, ET is a fundamental component in numerous studies based on the water balance framework (Pan et al., 2012; Zhang et al., 2012; Ukkola and Prentice, 2013; Oliveira et al., 2014; Wang et al., 2014b,a, 2015; Liu et al., 2016; Chen et al., 2020; Pascolini-Campbell et al., 2020; Wong et al., 2021; Ruhoff et al., 2022; Tan et al., 2022; Michailovsky et al., 2023; Unnisa et al., 2023; Xiong et al., 2023).

The water balance equation provides a physically grounded method to estimate ET by combining observed or modeled data on precipitation, river discharge, and terrestrial water storage change. Its basic formulation is expressed as:

$$ET = P - Q - \frac{dS}{dt} \qquad (1)$$

where P is the precipitation, Q denotes river discharge and $\frac{dS}{dt}$ is the change in water storage over time. Various studies, particularly those analyzing long-term periods (annual or more), considered the change in water storage ($\frac{dS}{dt}$) as negligible under the assumption that, over extended timescales, the inputs and outputs balance out, resulting in minimal net change in storage (Hobbins et al., 2001; Zhang et al., 2012; Hasenmueller and Criss, 2013; Xue et al., 2013). However, the validity of this assumption depends on the temporal scale, as well as the specific hydrological characteristics of the region under investigation (Zeng et al., 2012; Wu et al., 2019; Han et al., 2020).

Water balance basin analysis is particularly valuable for quantifying water availability, diagnosing the influence of climatic drivers on hydrological regimes, and supporting water resource management. Some studies have analyzed the water budget and the sources of imbalances (Pan et al., 2012; Wang et al., 2014b, 2015; Tan et al., 2022; Unnisa et al., 2023). Furthermore, it is commonly used to estimate ET itself. In this context, ET derived from the water balance serves to assess independent ET products and for conducting climate studies. As mentioned by Sheffield et al. (2009) and Oliveira et al. (2014), estimation of water balance components using ground-based measurements remains challenging due to uneven monitoring networks, expensive operations, and inadequate data openness and availability. To address these challenges, remote sensing offers a promising alternative by delivering spatially consistent estimates of terrestrial water cycle components across regional to global scales (Sheffield et al., 2009, 2018).

Satellite-derived data, particularly evapotranspiration and precipitation estimates, play an important role in optimizing the accuracy and applicability of the water balance equation. From the available products, the EUMETSAT CM SAF (Climate Monitoring Satellite Application Facility) datasets are distinguished by a high temporal resolution, long-term continuity, and explicit design for climate monitoring purposes. These products ensure temporal homogeneity and are particularly well-suited for hydrological and climatological studies. Their open-access availability also contributes to their value for regions with poor ground-based observations and research. In this context, two recent CM SAF products offer promising datasets for water balance analysis: the LANDFLUX Ed. 1 ET dataset (Moutier et al., 2024) and the GIRAFE v1 precipitation dataset (Konrad et al., 2025). The LANDFLUX dataset includes ET estimates at hourly, daily, and monthly resolutions over a 0.05°grid (approximately 5.5 km) covering the Meteosat disk (65°N–65°S and 65°W–65°E) for the 1983–2020 period. The GIRAFE dataset has global daily precipitation estimates, including sampling uncertainty and monthly means, at a spatial resolution of 1.0°, covering the period 2002–2022. Despite advancements in satellite sensors and retrieval techniques, closing the surface water budget using remote sensing data remains challenging. A recent paper by Zhang et al. (2023a) found substantial discrepancies between energy-balance-based ET (ET_{EB}) and water-balance-based ET (ET_{WB}) across 53 catchments in central-western Europe. These discrepancies were particularly pronounced in energy-limited catchments. At the annual scale, ET_{EB} showed weak agreement with ET_{WB} (correlation = 0.35), whereas at the monthly scale the correlation was stronger (r = 0.73). Such differences underscore the importance of basin characteristics, temporal aggregation, and input dataset choices in ET estimation. Given these challenges, it is essential to assess the suitability of CM SAF datasets for regional water balance applications. This study aims to: (i) evaluate the applicability of the CM SAF ET product in the context of water balance framework, (ii) compare water-balance-derived ET with independent remote sensing and reanalysis products, and (iii) assess water balance closure (i.e., the residual or imbalance) under various configurations of input datasets, including three distinct precipitation products, across a wide range of basin types. The analysis is conducted over 332 river basins across Europe. To provide a comprehensive evaluation of our dataset, we included a comparison with four widely used evapotranspiration products (LSA SAF, ERA5-Land, GLEAM, and GLDAS 2.1). Following the approach of Zhang et al. (2023a), we also investigate the influence of accounting for water storage changes and the role of basin characteristics, such as size and climate regime, on ET estimation accuracy. Furthermore, the choice of precipitation dataset, often a major source of uncertainty, is critically examined. Through this comprehensive evaluation, the

Deleted: 342

study provides insights into the strengths and limitations of CM SAF products for both scientific research and operational water resource management. It also contributes to the broader effort of integrating remote sensing data into hydrological assessments and climate monitoring frameworks.

2 Material and Method

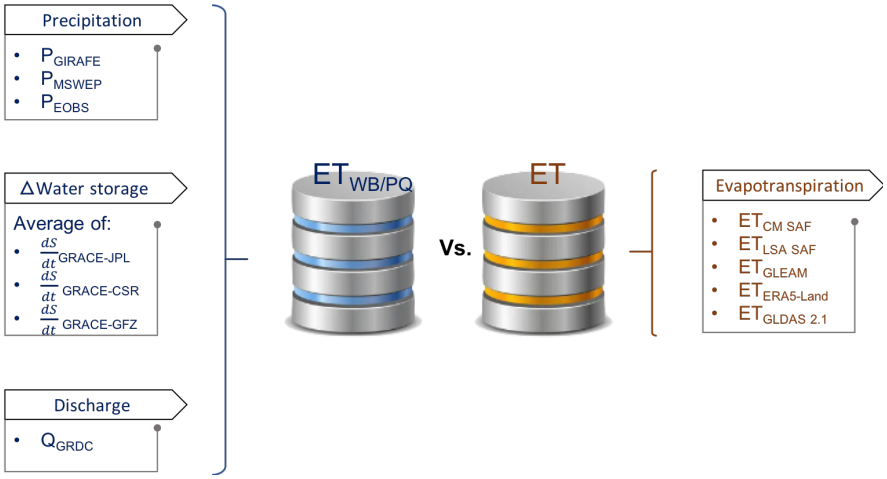


Figure 1: Schematic representation of data sources and methodologies used for estimating and comparing evapotranspiration based on the water balance method including (ET_{WB}) or not (ET_{PQ}) the change in water storage with remote sensing-based/reanalysis products.

Figure 1 shows the approach which is used, at each selected basin, in comparing evapotranspiration (ET) derived from various independent datasets (satellite-based approach, reanalysis datasets) and those estimated using the water balance equation with (ET_{WB}) and without (ET_{PQ}) considering the change in water storage (assumed to be zero in the latter). On the left, the components of the water balance equation are presented, including precipitation (P, see Section 2.3) from various sources (GIRAFE, MSWEP, and E-OBS), water storage changes ($\frac{dS}{dt}$) derived from GRACE data (averaged across JPL, CSR, and GFZ solutions, see Section 2.4), and discharge (Q) from the GRDC dataset (see Section 2.1). On the right, in addition to the CM SAF ET product, other ET products were also considered in the analysis for comparison purposes. Those products, selected for their widespread use within the scientific community, are the LSA SAF, GLEAM, ERA5-Land, and GLDAS 2.1 ET products. The imbalance will also be analyzed by calculating the difference between water balance-derived ET and independent products:

$$\varepsilon = \frac{\sum_{i=0}^{n} P_i - Q_i - \frac{dS_i}{dt} - ET_{P_i}}{N} = \frac{\sum_{i=0}^{n} ET_{WB_i} - ET_{P_i}}{N} \quad (2)$$

Field Code Changed
Deleted: Figure 1
Formatted: Font: Not Bold
Deleted:
Formatted: Font: (Default) Arial, 11 pt
Deleted: (ET_{PQ})
Field Code Changed
Formatted: Font: Times New Roman, 12 pt

where N is the number of estimations, ET_{P_i} corresponds to evapotranspiration values for various independent products (CM SAF, LSA SAF, GLEAM, GLDAS 2.1 and ERA5-Land) and ET_{WB_i} represents the evapotranspiration values calculated using equation 1 (ET_{WB} can be replaced by ET_{PQ}). Furthermore, specific analyses will be conducted based on basin size (larger or smaller than 90 000 km^2; see Section2.4) and climate regime (water- or energy-limited basins; see Section 2.5).

A detailed description of the used datasets is provided in Table 1 and the following subsections. From their native spatial resolution, all datasets have been remapped to a common spatial (0.25°) and temporal (monthly or yearly) resolutions. This ensures consistency across datasets and allows direct comparisons over different timescales.

Table 1: Overview of the datasets used in this study, including their names, temporal and spatial resolution (as used in this study), and periods of availability.

Products				
	Dataset name	**Temporal resolution**	**Native spatial resolution**	**Period**
Evapotranspiration	LANDFLUX Ed.1 (CM SAF)	Monthly	0.05°	1983-2020
	LSA SAF v3 (DR)	Daily	0.05°	2004-2020
	GLEAM v4.2a	Monthly	0.1°	1980-2023
	GLDAS 2.1	Monthly	0.25°	2000-present
	ERA5-Land	Monthly	0.1°	1950-present
Precipitation	GIRAFE (CM SAF)	Monthly	1°	2002-2022
	E-OBS v23.1e	Daily	0.25°	1950-present
	MSWEP v2.8	Monthly	0.1°	1979-present
Water storage	GRACE v6.3 CSR RL06	Monthly	1°	2002-~present
	GRACE v6.3 JPL RL06	Monthly	1°	2002-~present
	GRACE v6.3 GFZ RL06	Monthly	1°	2002-~present
Discharge	GRDC	Monthly	station	1806- present

Deleted: (CDR)

Commented [MOU1]: Not 2022

2.1 Discharge Global Runoff Data Center (GRDC)

The monthly discharges data were collected from the Global Runoff Data Center (GRDC). Established in 1988, the GRDC maintains the most comprehensive and quality-controlled collection of river discharge data worldwide. It archives river discharge data up to 216 years old at both daily and monthly scales for more than 10 800 stations from 160 countries. Discharge values are originally provided in cubic meters per second ($m^3\ s^{-1}$) and have been converted to mm month^{-1} (Q (mm month^{-1}) = Q ($m^3\ s^{-1}$) × 86400000 × number of days in month /Area(m^2)). The GRDC river discharge database includes essential metadata such as station coordinates, station and river names, upstream catchment area, elevation, long-term mean discharge, and a shapefile delineating the contributing basin. The shapefile of the basin is used to spatially aggregate other water balance components from gridded datasets. By averaging these variables over each basin, consistent input values are obtained for application in the water balance equation.

For this study, we selected stations that met the following criteria: (i) a drainage area greater than 1 000 km^2, (ii) a minimum of five consecutive years of data, and (iii) availability of valid data from all other datasets covering 100% of the basin surface for each station. Based on these criteria, 332 stations across Europe (Figure 2) were used for the analysis.

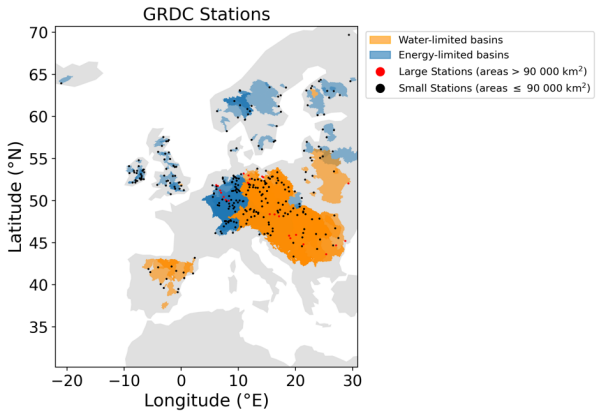


Figure 2: Spatial distribution of the 332 river discharge stations selected from the Global Runoff Data Centre (GRDC) across Europe. Each dot represents a station, with red indicating large basins (area > 90 000 km^2) and black indicating small basins (area ≤ 90 000 km^2). The associated basin shapes are also shown, colored according to their aridity index (AI): blue for energy-limited basins (AI ≤ 1) and orange for water-limited basins (AI > 1).

Deleted: ¶

Formatted: Font: Not Bold

Deleted: for the station

Deleted: 342

Formatted: English (UK)

Commented [MOU2]: 90000 instead of 9000

Deleted: 342

To account for variations in basin size and climate regime, we performed specific analyses based on these factors. First, since GRACE provides more reliable estimates $\frac{dS}{dt}$ for basins larger than 90 000 km^2 (Tangdamrongsub et al. (2021), see Section 2.4 for details), we categorized the basins into two groups: 25 large basins (area >= 90 000 km^2) and 307 smaller basins. Additionally, based on an estimation of the aridity index (AI) (see Section 2.5), the basins were classified into two climate categories: 94 "energy-limited" (wet; AI ≤1) and 238 "water-limited" (dry; AI>1) basins (see details in Section 2.5).

For visualization purposes, among the 25 stations located in basins larger than 90 000 km^2 (across seven rivers), the largest station from each river was selected to provide a representative sample of the major basins. This results in a subset of seven stations (see Table 2 in the Appendix), hereafter referred to as the six largest river stations for simplicity.

Deleted: 26
Deleted: 316
Deleted: 95
Deleted: 247
Deleted: 26
Deleted: seven largest

2.2 Actual evapotranspiration

2.2.1 CM SAF LANDFLUX Ed. 1 dataset

The CM SAF Surface Radiation and Fluxes - Edition 1 (LANDFLUX Ed. 1; Moutier et al. (2024)) dataset provides nearly 40 years (1983–2020) of parameters depicting the surface states and radiation fluxes, including the Surface Radiation Balance (SRB), Cloud Fractional Cover (CFC), Land Surface Temperature (LST), Evapotranspiration (ET), and Latent (LE) and Sensible (H) Heat Fluxes. Retrievals are based on two sensors aboard the Meteosat suite of geostationary satellites: the Meteosat Visible and InfraRed Imager (MVIRI) and the Spinning Enhanced Visible and InfraRed Imager (SEVIRI).

The parameter ET is calculated using an adapted version of the methodology developed by the Land surface Land Surface Analysis SAF (Barrios et al. (2024)), based on the Hydrology Tiled ECMWF Scheme for Surface Exchanges over Land (H-TESSEL; van den Hurk et al., 2000; Balsamo et al., 2009). This adaptation allows the use of both satellite-based data and numerical weather prediction (NWP) model outputs as forcing inputs. A full description of the methodology is described in Moutier et al. (2023b).

LANDFLUX Ed. 1 data are available in hourly, daily, and monthly temporal resolutions, as well as monthly diurnal cycle composites. The dataset covers the ±65° longitude and ±65° latitude region on a 0.05°× 0.05°regular grid. In the following, only monthly data will be used. To simplify the reading, this dataset will hereafter be referred to as the CM SAF evapotranspiration product.

2.2.2 LSA SAF dataset

The LSA SAF ET dataset is a satellite-based product based on SEVIRI observations (Barrios et al., 2024). The data record (DR) demonstrator provides 30-minute and daily evapotranspiration estimates between 2004 and 2020 at a 0.05°× 0.05°spatial resolution over the Meteosat coverage area. As discussed in Section 2.2.1, the LSA SAF and the CM SAF

share a similar core model, with key differences arising from the choice of input data. For instance, while CM SAF uses ERA5 as input to estimate the soil moisture, the LSA SAF approach relies on the H SAF soil moisture dataset. Other differences can be seen in the choice of the stomatal resistance values and the retrieval of the leaf area index (LAI) for each tile. For instance, an inversion matrix approach is used in the LSA SAF (Ghilain et al., 2011, 2012) while a LUT approach is adopted in the CM SAF (Moutier et al., 2023b).

2.2.3 ERA5-Land dataset

ERA5 dataset (Hersbach et al., 2019) is the fifth generation of global atmospheric reanalysis generated by the ECMWF. The underlying land surface model used is H-TESSEL. ERA5 ET is derived from the ERA5 atmospheric reanalysis, which includes coupled atmosphere–land interactions. All global atmospheric, oceanic and land surface fields are available at an hourly time step with a spatial resolution of 0.25°(~32km) covering the period from January 1950 to present, thus the full extend of the record. ERA5-Land (Muñoz Sabater et al., 2021) is a high-resolution land component of the ERA5 atmospheric reanalysis, providing data at a spatial resolution of 0.1°× 0.1°. ERA5-Land is generated using the same land surface model but forced with downscaled meteorological variables from ERA5 without the direct feedback of atmospheric coupling. This later demonstrated its performance in simulating the evapotranspiration in offline experiments (Muñoz Sabater et al., 2021, and reference therein).

2.2.4 GLDAS dataset

Global Land Data Assimilation System version 2 (GLDAS; Rodell et al., 2004) is a new generation of reanalysis developed jointly by the National Aeronautics and Space Administration (NASA) Goddard Space Flight Center (GSFC) and National Center for Environmental Prediction (NCEP). GLDAS, which has been streamlined and parallelized by the Land Information System (LIS; Kumar et al., 2006), generates land surface products by using various offline (not coupled to the atmosphere) land surface models (LSM) and ingesting satellite- and ground-based observational datasets (Rodell et al., 2004). Details about the forcing data and description of the model are available on http://disc.Sci.GSFC.NASA.Gov/Hydrology. Currently, GLDAS has three components: GLDAS-2.0 (1948-2014), GLDAS-2.1 (2000- Present), and GLDAS-2.2 (Feb 2003-Present). Beyond their differences in term of forcing or data assimilation source, the choice of the product has been dictated by the period covered to validate the CM SAF product. We have selected the GLDAS-2.1 product with a monthly temporal resolution. GLDAS-2.1 data have been simulated by the Noah 3.6 Model in Land Information System Version 7 with a spatial resolution of 0.25°. This simulation was forced with National Oceanic and Atmospheric Administration (NOAA)/Global Data Assimilation System (GDAS) atmospheric analysis fields (Derber et al., 1991), the disaggregated Global Precipitation Climatology Project (GPCP) V1.3 Daily Analysis precipitation fields (Adler et al., 2003; Huffman et al., 2001), and the Air Force Weather Agency's AGRicultural METeorological modeling system (AGRMET) radiation fields. The simulation was only used with GDAS and GPCP from January 2000 to February 2001, followed by the addition of AGRMET from March 1, 2001 onwards.

2.2.5 GLEAM dataset

The Global Land Evaporation Amsterdam Model (GLEAM; Miralles et al., 2011; Martens et al., 2017; Miralles et al., 2024) is a remote sensing-based model allowing the estimation of the terrestrial evapotranspiration components, including transpiration, bare soil evaporation, interception loss, sublimation, as well as root-zone soil moisture. To account for random errors in the forcing data and other processes not explicitly represented in the model, such as irrigation, GLEAM assimilates microwave soil moisture (SMs) and/or backscatter observations into the soil profile. Interception loss is calculated separately using the approach from Zhong et al. (2022). Potential evapotranspiration is calculated using the Priestley-Taylor (PT) equation (Priestley and Taylor, 1972) and, the actual evapotranspiration is derived by including stress factors such as soil moisture states and vegetation physiological characteristics. Key features of this model include the integration of microwave-derived soil moisture, land surface temperature, and vegetation density, along with a detailed parameterization of rainfall interception loss. The GLEAM dataset is available globally at daily, monthly and yearly temporal resolutions, with a spatial resolution of 0.1°, covering the period 1980 to 2023. In this use case, we use GLEAM V4.2a product at a monthly temporal resolution with a spatial resolution of 0.1°.

Deleted: —

2.2.6 Intercomparison of evapotranspiration products

Figure 3 displays the probability density functions (PDFs) of monthly ET estimates for all selected basins from the five datasets reggrided at 0.25°: CM SAF, LSA SAF, GLEAM, ERA5-Land, and GLDAS-2.1. The CM SAF and LSA SAF distributions exhibit a strong peak at lower ET values, indicating a higher frequency of low ET estimates. In particular, mode values are of 3 mm month^{-1} for CM SAF and 6 mm month^{-1} for LSA SAF datasets whereas other datasets show higher values ranging from 8 (GLDAS) to 12 (GLEAM) mm month^{-1}. Furthermore, CM SAF presents a lower median value of 21 mm month^{-1}, compared to 36 ± 2 mm month^{-1} for other datasets. This suggests that CM SAF dataset tends to produce slightly lower values. This finding is aligned with the results reported in Moutier et al. (2023a). This comparison underscores the variability in ET estimates across different datasets and emphasizes the importance of understanding the underlying model assumptions and forcing data when analyzing water balance components.

Deleted: 5

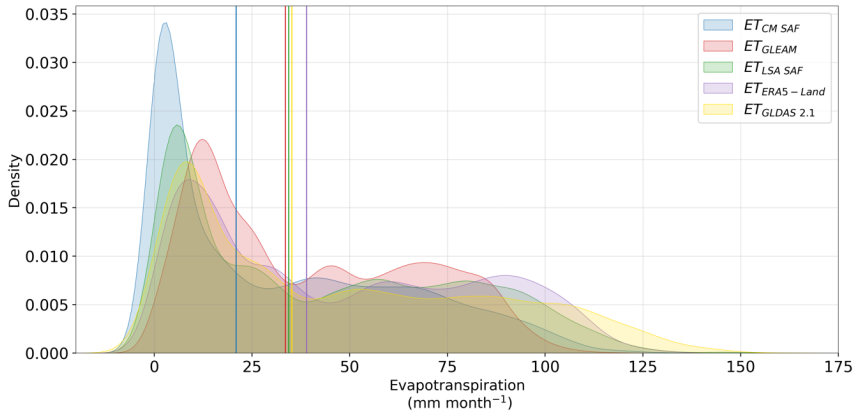


Figure 3: Probability density functions (PDFs) of monthly evapotranspiration (ET) estimates for all selected basins, over the period 2004-2020, from the five regridded (0.25°) datasets: CM SAF (blue), LSA SAF (red), GLEAM (green), ERA5-Land (purple), and GLDAS-2.1 (yellow). Vertical solid lines indicate the median ET values for each dataset.

2.3 Precipitation

2.3.1 E-OBS

E-OBS v23.1e is an ensemble gridded dataset of surface climate observations (precipitation sum, mean sea level pressure, mean wind speed, mean relative humidity, global radiation and mean, minimum and maximum temperature) at daily resolution for Europe (25°N-71.5°N x 25°W-45°E; Cornes et al., 2018). These variables are available on a 0.1°and 0.25° land-only regular grid, offering detailed spatial resolution. The E-OBS gridded data set is derived through interpolation of observations (see details in Haylock et al., 2008; Cornes et al., 2018) from a dense network of meteorological stations data collated by the ECA&D initiative (European Climate Assessment and Data; Klein Tank et al., 2002; Klok and Klein Tank, 2009). Since the initial construction of E-OBS by Haylock et al. (2008), the number of stations has drastically changed, with approximately 2,500 to approximately 9000 stations in the case of precipitation. The uncertainty of the product is not homogeneous over the grid as the product uncertainty increases in data-sparse regions. However, the robustness and reliability of E-OBS are further ensured through rigorous quality control measures and the inclusion of a wide range of meteorological stations, making it suitable for this use case. Before extracting the basin shape, monthly data was obtained by accumulating the daily product.

2.3.2 GIRAFE

The CM SAF has recently released its first global precipitation climate data record: the Global Interpolated RAinFall Estimation version 1 (GIRAFE v1; Konrad et al., 2025). It provides daily accumulated precipitation and respective sampling uncertainty, as well as monthly mean values, covering the period 2002–2022 at a 1.0°spatial resolution. Estimations are derived from a combination of passive microwave (PMW) observations onboard polar orbiting satellites and infrared (IR) observations onboard geostationary satellites distributed around the equator (Geo-Ring). Originally developed for the Megha-Tropiques satellite mission, the Tropical Amount of Precipitation with an Estimate of ERrors (TAPEER; Chambon et al., 2013; Roca et al., 2018) approach has been adapted to merge PMW and IR data as well as the estimation of the daily sampling uncertainty. In the 55°S–55°N latitude range, PMW-derived precipitation rates are merged with IR-based precipitation occurrence to compute daily precipitation accumulations. Beyond the 55°S–55°N latitude range, the IR Geo-Ring pixel distortion is too high, and daily accumulation estimation exclusively relies on PMW observations. Thus, the GIRAFE sampling uncertainty, being estimated from IR observations, is only available in the 55°N/S latitude region. Please note that a full description of the method can be found in (Konrad et al., 2025; Niedorf et al., 2024). In the framework of this study, only monthly data has been extracted and remapped to a common 0.25°grid using the *"remapcon"* function of the Climate Data Operators (CDO) tool (Schulzweida, 2023).

2.3.3 MSWEP

The Multi-Source Weighted-Ensemble Precipitation v2.8 (MSWEP; Beck et al., 2019) dataset is a high-resolution, global precipitation dataset that combines information from multiple sources, including gauge observations, satellite estimates, and reanalysis data. The dataset is unique in being able to incorporate daily gauge observations and correct for gauge reporting times, which minimizes temporal mismatches between the different data sources (see Beck et al. (2017, 2019) for more information). MSWEP v2.8 has a 3-hourly, daily, and monthly temporal resolution and a spatial resolution of 0.1° in three variants: "NRT", "Past_nogauge", and "Past". The "NRT" is the near real-time extension (with latency of 3 hours). The "Past" and "Past_nogauge"" are the historical satellite-reanalyses merged with and without gauge corrections, respectively, for the period 1979–2020. In the framework of this study, we have used the 'Past' version as it includes all the data sources of "Past_nogauge" but also includes daily gauge corrections. This version provides the optimal precipitation estimates by merging gauge data and is suitable for hydrological and climatological purposes.

2.3.4 Intercomparison of precipitation products

Figure 4 displays the probability density functions (PDFs) of monthly precipitation (P) estimates for all selected basins from the three datasets regridded at 0.25°: GIRAFE, MSWEP and E-OBS. The GIRAFE distributions exhibit lower precipitation values as compared to MSWEP and E-OBS products. Indeed, the median value for GIRAFE is of 31 mm month^{-1} while medians of 65 and 62 mm month^{-1} are observed for MSWEP and E-OBS, respectively. This observation aligns with the findings by Konrad et al. (2025), where GIRAFE tends to produce lower values over Europe. This underestimation could be partially related to the inadequate detection of precipitation over snow and ice surfaces (Konrad et al., 2025) and a general tendency toward lower daily precipitation (Konrad et al., 2024). To further illustrate this, an extended version of

Formatted: Justified

Deleted: via a bilinear interpolation

Formatted: Font: (Default) Arial, 11 pt

Formatted: Font: Italic

Formatted: Font: (Default) Arial, 11 pt, Italic

Formatted: Font: Italic

Formatted: Font: (Default) Arial, 11 pt

Formatted: Font: 11 pt

Formatted: Font: (Default) Arial

Deleted: .

Deleted: 2

Deleted: 1

Formatted: Font: (Default) Arial, 11 pt

the PDF analysis is provided in Appendix Figure 19, which shows how applying different thresholds on the number of snow days affects the distribution of GIRAFE precipitation values.

Formatted: Font: (Default) Arial, 11 pt, Not Bold

Formatted: Font: (Default) Arial, 11 pt

Formatted: Font: (Default) Arial, 11 pt

Formatted: English (US)

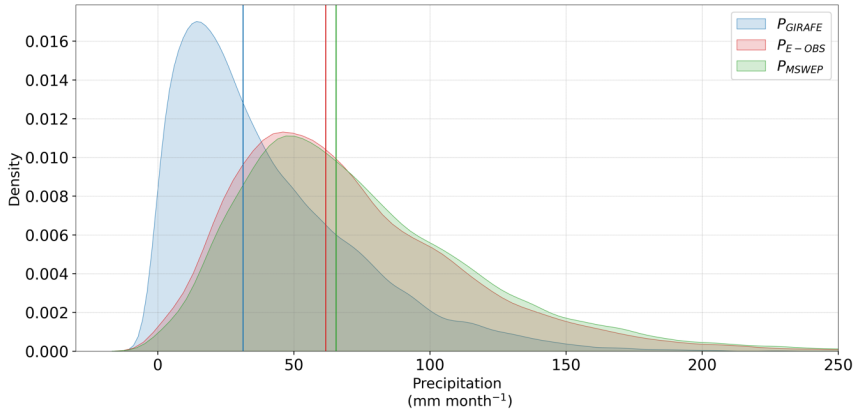


Figure 4: Probability density functions (PDFs) of monthly precipitation (P) estimates for all selected basins, over the period 2004-2020, from the three regridded (0.25°) datasets: GIRAFE (blue), MSWEP (red) and E-OBS (green). Vertical solid lines indicate the median precipitation values for each dataset.

2.4 Water storage

For each basin, the changes in total water storage ($\frac{dS}{dt}$ at monthly timescale) are estimated using data from NASA Gravity Recovery And Climate Experiment (GRACE) and its follow-up mission, the Gravity Recovery and Climate Experiment Follow-on (GRACE-FO) data (Landerer and Swenson, 2012). These missions provide estimates of terrestrial water storage by detecting variations in Earth's gravity field caused by mass redistribution, primarily associated with changes in water storage. The GRACE-derived total water storage is available for the period 2002–near present at a spatial resolution of 1°based on three RL06 solutions provided by major processing centers: GFZ (GeoForschungsZentrum; Landerer, 2021b, 2023a), CSR (Center for Space Research, University of Texas; Landerer, 2021a, 2024), and NASA JPL (Jet Propulsion Laboratory; Landerer, 2021c, 2023b). The estimation of water storage changes follows a well-established approach (Swann and Koven, 2017), where the monthly variation in water storage is computed using a centered finite difference of the monthly total water storage values:

$$\frac{dS}{dt} \approx \frac{S_{t+1} - S_{t-1}}{2\Delta t} \quad (3)$$

where Δt represents a time step of one month, and S_{t+1} and S_{t-1} correspond to the monthly mean water storage in the basin for the following and preceding months, respectively.

Finally, we have used a simple average of the three different GRACE spherical harmonic solutions (GFZ, CSR, and JPL) to reduce uncertainties associated with individual processing strategies and enhance the robustness of total water storage estimates (Sakumura et al., 2014). As described by Tangdamrongsub et al. (2021), the sensitivity of GRACE to changes in terrestrial water storage is constrained by the design of its satellites and orbital configuration. GRACE is more accurate while detecting variations over relatively large areas (e.g., >90 000 km^2) or mass changes exceeding $\simeq$1 Gton, which corresponds to approximately 1 cm of equivalent water height over 3°×3°grid cell. This coarse spatial resolution limits its applications for detailed $\frac{dS}{dt}$ assessments in smaller river basins. Nevertheless, numerous studies include smaller basins into their analyses (e.g., Zeng et al., 2012; Wang et al., 2014b; Xiong et al., 2023; Zhang et al., 2023a), demonstrating that GRACE data can still be valuable at finer scales. Therefore, this limitation has been used as a threshold to categorize basins into different groups for further analysis (see Section 2.1).

Please note that the use of a centered finite difference approach ideally requires a continuous time series. However, the original GRACE mission terminated in June 2017, while its successor, GRACE-FO, only began data acquisition in June 2018, resulting in an 11-month gap between July 2017 and May 2018. Furthermore, the GRACE mission itself contains additional missing months, and certain months were excluded due to poor temporal centering. A data file spanning from 20 March 2012 to 20 April 2012 is not considered to be centered on a specific month and was therefore excluded from the analysis. Although various gapfilling methods have been proposed to address these discontinuities (Zhang et al., 2025), the implementation of such techniques was beyond the scope of this study.

2.4.1 Intercomparison of change in water storage products

Figure 5 shows the probability density functions (PDFs) of monthly changes in water storage estimates across all selected basins, based on the three GRACE solutions, JPL, GFZ, and CSR, regridded to 0.25°. The spread of the distributions suggests variability between the datasets, suggesting differences in how each solution captures changes in water storage. Despite this variability, the distributions are generally centered around zero, indicating, over the period 2004-2020, no pluriannual accumulation or depletion of water at the basin scale. Nevertheless, the variability implies the presence of fluctuations at finer temporal or spatial resolutions.

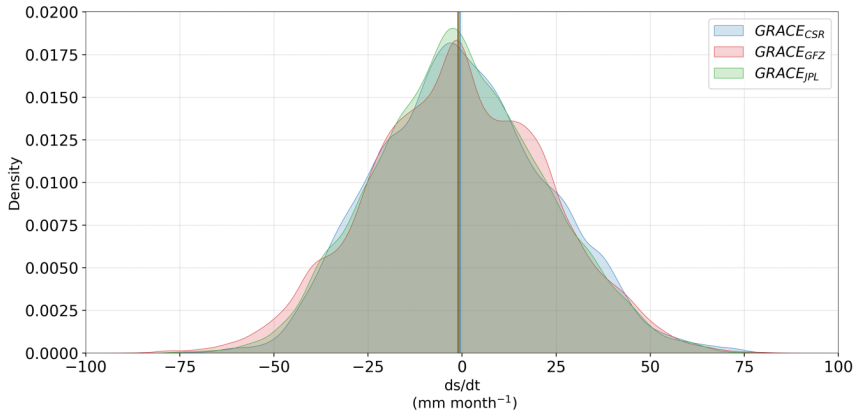


Figure 5: Probability density functions (PDFs) of monthly water storage ($\frac{dS}{dt}$; mm month^{-1}) estimates for all selected basins, over the period 2004-2020, from the three sources: GRACE-CSR (blue), GRACE-JPL (red) and GRACE-GFZ (green). Vertical solid lines indicate the median values for each PDFs.

2.5 Aridity index

The aridity index (AI), here defined as the ratio of annual reference evapotranspiration ET_{ref} to annual precipitation (though the potential ET is sometimes used as well; Zarch et al., 2015; Kukal and Irmak, 2016, and references therein), is a key indicator of climate dryness in a given region. An AI of 0 corresponds to extremely arid areas with no precipitation, while values close to 1 indicate humid environments where most precipitation is returned to the atmosphere through evapotranspiration. Values greater than 1 represent areas where precipitation exceeds evapotranspiration. To distinguish the impact of climate on water balances, we followed the Palmer et al. (2010) and Zhang et al. (2023a) approaches by classifying basins as either water-limited (AI > 1) or energy-limited (AI ≤ 1).

In the framework of this study, we derive several AI thanks to the LSA SAF ET_{ref} daily product (Trigo et al. (2011); Paredes et al. (2021)), and the three precipitation products (see section 2.3). ET_{ref} LSA SAF and precipitation data were initially remapped to a common 0.25°grid using the *"remapcon"* function of the CDO tool. Subsequently, the daily/monthly mean values were calculated for each basin and aggregated to annual timescales, considering only years with a full set of 12 valid months. Yearly ET_{ref} and P values have been averaged to finally obtain three AI estimates: AI_{GIRAFE}, $AI_{E\text{-}OBS}$, and AI_{MSWEP}. A fourth AI was computed as the median of the three individual estimates and named $AI_{Average}$.

Commented [MOU3]: Inversion in text.

Deleted: bilinear interpolation

2.5.1 Intercomparison of AI products

Figure 6 presents these four aridity indices' probability density functions (PDFs). As expected, the lower precipitation values from GIRAFE (see Section 2.3.2) result in a higher AI as compared to other datasets. The median value for AI_{GIRAFE} is 1.5, whereas median values for AI_{MSWEP} and $AI_{E\text{-}OBS}$ are 0.8 and 0.84, respectively. By using $AI_{Average}$ for the classification, 94 basins were identified as water-limited and 238 as energy-limited.

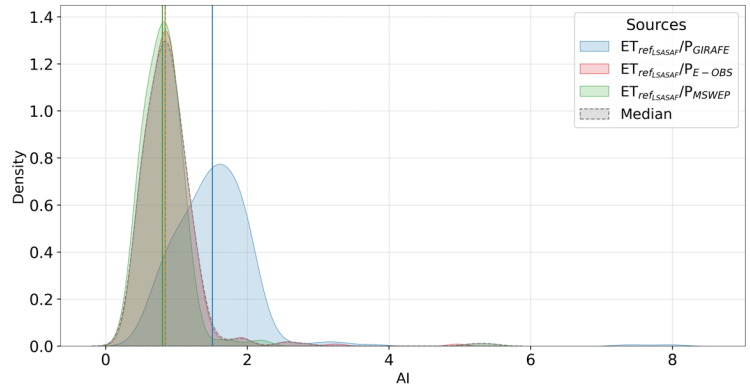


Figure 6: Probability density functions (PDFs) of the aridity index (AI), computed as the ratio of reference evapotranspiration from LSA SAF ($ET_{ref_{LSA\ SAF}}$) to precipitation (P), for all selected basins over the period 2004–2020. The analysis is based on three regridded (0.25°) precipitation datasets: GIRAFE (blue), E-OBS (red), and MSWEP (green). The grey-shaded area represents the distribution of median AI values derived from the three individual datasets. Vertical solid lines indicate the median AI for each dataset.

2.6 Statistical metrics

The evaluation is made in terms of bias, mean absolute difference (MAD) and unbiased root mean square difference (uRMSD):

$$Bias = \frac{\sum_{i=0}^{n} ET_{P_i} - ET_{WB_i}}{N} \quad (4)$$

Deleted: 1
Deleted: 5
Deleted: 247
Deleted: green
Deleted: red
Formatted
Deleted: 4

$$MAD = \frac{\sum_{i=0}^{n} |ET_{P_i} - ET_{WB_i}|}{N} \quad (5)$$

$$uRMSD = \left[(RMSD)^2 - (Bias)^2\right]^{1/2} \quad (6)$$

with,

$$RMSD = \left[\frac{\sum_{i=0}^{n} ET_{P_i} - ET_{WB_i}}{N}\right]^{1/2} \quad (7)$$

where N is the number of estimations, ET_{P_i} corresponds to evapotranspiration values for various independent products (CM SAF, LSA SAF, GLEAM, GLDAS 2.1 and ERA5-Land) and ET_{WB_i} represents the evapotranspiration values calculated using equation 1 (ET_{WB} can be replaced by ET_{PQ}).

Deleted: 5
Formatted
Deleted: 6
Formatted
Formatted
Deleted: 7

3 Results & Discussion

3.1 Specific example for the Düsseldorf station (Rhine River)

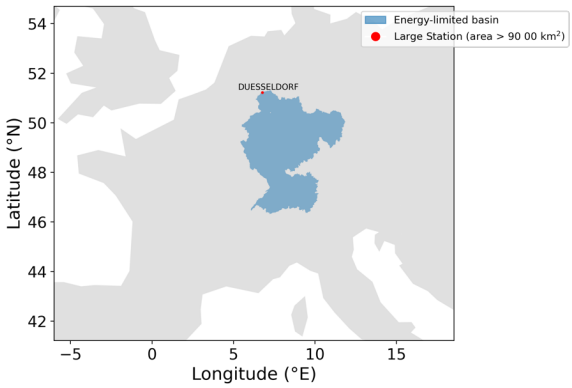


Figure 7: Location of the GRDC station Düsseldorf (red dot) and its associated river basin (blue polygon) derived from GRDC shapefiles (area of 147680 km^2). The basin, part of the Rhine River system, is classified as large (area > 90 000 km^2) and energy-limited (AI ≤ 1).

Analyzing a single basin provides a valuable opportunity to examine key hydrological components individually, gaining information about their magnitudes and variability. This approach not only enhances the understanding of each component separately but also helps to identify inconsistencies between different data sources or estimation methods. For instance, discrepancies between water balance-derived evapotranspiration (ET_{WB} and ET_{PQ}) and satellite-based or reanalysis estimates can indicate biases in precipitation, discharge, or water storage change components. In the same way, variations among precipitation datasets or GRACE-derived storage changes from different processing centers can reveal uncertainties related in data source/processing choices. This analysis clarifies how precipitation, discharge, and change in water storage are combined to estimate the overall water balance evapotranspiration, ultimately improving confidence in remote sensing based and reanalysis product-based evapotranspiration estimates.

The Rhine River (Düsseldorf GRDC station, basin area of 147680 km^2; Figure 7), holds significant importance due to its hydrological significance, economic impact, and role in climate studies, offering insights into water management challenges and climate change impacts in a major European river basin (Görgen et al., 2010; Bosshard et al., 2014; Stahl et al., 2022).

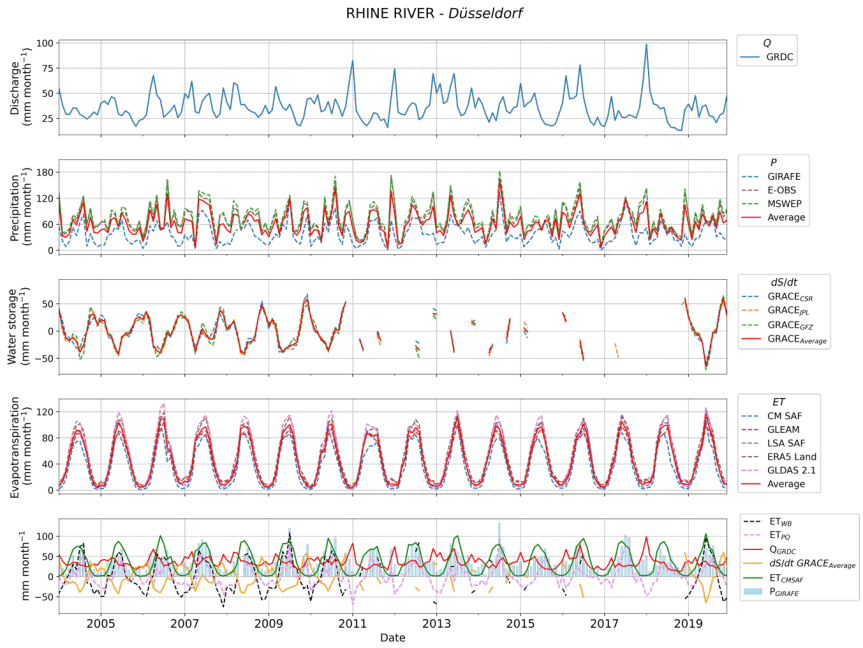


Figure 8: Monthly time-series of key hydrological components for the Rhine River at Düsseldorf (2003–2020). The panels display, from top to bottom: (1) observed discharge (Q_{GRDC}); (2) precipitation (P) from multiple datasets (GIRAFE, E-OBS, MSWEP), including their mean ($P_{Average}$, red solid line); (3) water storage change ($\frac{dS}{dt}$) estimated from GRACE solutions (CSR, JPL, GFZ), with their ensemble mean (red solid line); (4) evapotranspiration (ET) estimates from different datasets (CM SAF, LSA SAF, GLEAM, ERA5-Land, GLDAS 2.1), including their mean (red solid line); and (5) water balance-derived evapotranspiration with (ET_{WB}) and without (ET_{PQ}) including the water storage component, along with the monthly averages of key hydrological variables used in their computation: discharge (Q_{GRDC}; red solid line), mean water storage ($\frac{dS}{dt}GRACE_{Average}$; yellow solid line), GIRAFE precipitation (P_{GIRAFE}, bar-plot in blue) and the CM SAF evapotranspiration ($ET_{CM\ SAF}$).

Figure 8 presents the monthly time series of key hydrological components for the Rhine River at Düsseldorf from 2003 to 2020. The first panel illustrates the observed discharge (Q_{GRDC}), which exhibits variability, with peak values exceeding 75 mm month^{-1} during high-flow periods and minimum values around 20 mm month^{-1} (5th percentile (perc) = 17.9 mm month^{-1}). The second panel shows monthly precipitation estimates from three datasets (GIRAFE, E-OBS, MSWEP), with the mean ($P_{Average}$). Precipitations exhibit substantial month-to-month variability, with 5th and 95th percentile of the distribution being 8 and 135 mm month^{-1} (including all products), respectively. The third panel presents the change in terrestrial water storage ($\frac{dS}{dt}$), estimated from GRACE solutions (CSR, JPL, GFZ) and averaged. Most variations are ranged between ±45 mm month^{-1} (more than ≃90% of the data) with minimal and maximal values obtained in June (≃-32 mm month^{-1}) and November (≃38 mm month^{-1}). The estimated change

Deleted: mean
Deleted: A
Deleted: verage
Deleted: mean
Deleted: from all independent datasets
Deleted: Average
Formatted: Subscript
Deleted: ¶
Deleted: 6
Deleted: 6
Deleted: 6

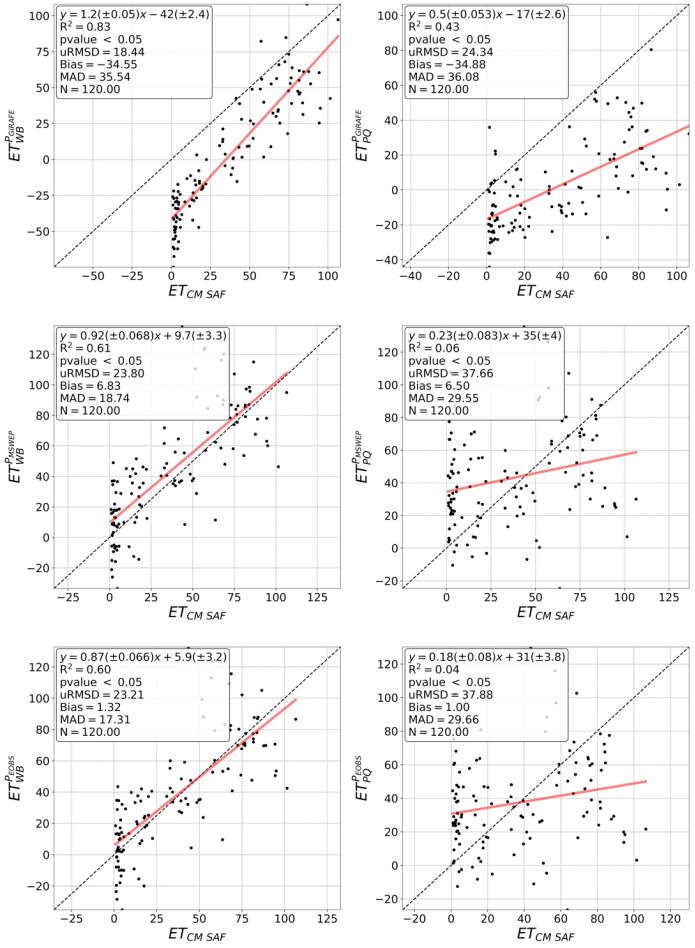


Figure 9: Scatter plots comparing monthly evapotranspiration estimates from the CM SAF product ($ET_{CM\ SAF}$ in mm month^{-1}) with water balance-derived evapotranspiration estimates with (ET_{WB} in mm month^{-1}; left panels) and without (ET_{PQ} in mm month^{-1}; right panels) considering $\frac{dS}{dt}$ for all basins. GIRAFE (top panels), MSWEP (middle panels), and E-OBS (bottom panels) precipitation products are considered for the calculation of ET_{WB} and ET_{PQ}. The dashed black line represents the 1:1 relationship. Statistical metrics, including the coefficient of determination (R^2), unbiased root mean square deviation (uRMSD), bias, mean absolute deviation (MAD), the number of observations (N), and the linear relationships, are provided in each panel.

in water storage varies according to the GRACE solution. Across the three datasets, the average maximum difference is of 10 mm month^{-1}, while the standard deviation is on average of 5.5 mm month^{-1}. For all methods, several gaps are observed between 2011 and 2018. As explained in the section 2.4, missing values are primarily due to data gaps between the GRACE and GRACE-FO missions, as well as the exclusion of months with missing or poorly centered observations, which are required for the finite difference approach. The fourth panel presents evapotranspiration (ET) estimates from various datasets (CM SAF, LSA SAF, GLEAM, ERA5-Land, GLDAS 2.1). A clear seasonal pattern is observed, with the maximum values of about 110 mm month^{-1} in summer and the minimum values of about 0 mm month^{-1} in winter. While the ET values tend to be in the same order of magnitude, the highest differences are observed in summer. The final panel shows precipitation, discharge, GRACE-derived storage change, and evapotranspiration. Two estimates of water balance-derived evapotranspiration are included: ET_{WB} and ET_{PQ}. These comparisons provide insights into potential discrepancies between datasets and the water balance approach.

With a specific focus on evapotranspiration, we observe similar temporal dynamics between ET_{WB} and $ET_{CM\ SAF}$, while ET_{PQ} shows a less consistent pattern. However, ET_{WB} presents significant data gaps. This similarity in the temporal courses is confirmed by the 1:1 comparison in Figure 9, where $ET_{CM\ SAF}$ is compared with ET_{WB} and ET_{PQ}, based on GIRAFE (ET_{WB}^{GIRAFE} and ET_{PQ}^{GIRAFE}), MSWEP (ET_{WB}^{MSWEP} and ET_{PQ}^{MSWEP}) and E-OBS ($ET_{WB}^{E\text{-}OBS}$ and $ET_{PQ}^{E\text{-}OBS}$) products. When $\frac{dS}{dt}$ is not considered, bias values are between 1 ($ET_{PQ}^{E\text{-}OBS}$) and -34.9 mm month^{-1} (ET_{PQ}^{GIRAFE}), but the correlation is weak with R^2 values between 0.04 ($ET_{PQ}^{E\text{-}OBS}$) and 0.43 (ET_{PQ}^{GIRAFE}). The availability of GRACE data is a limiting factor for the ET_{WB} timeseries. The analysis presented here uses a consistent subset of common match-ups across datasets to ensure statistical comparability. However, when this constrain is relaxed, the number of valid matchups increases significantly, by a factor of 1.6 to 3.2. This illustrates trade-off between statistical robustness and data availability when incorporating $\frac{dS}{dt}$. Conversely, when $\frac{dS}{dt}$ is included, bias values remain similar, but stronger relationships (≥0.57) are observed, and both the uRMSD and MAD tend to decrease. Only the MAD for GIRAFE remain nearly unchanged. Notably, the best R^2 is obtained with the GIRAFE (CM SAF) dataset with a value of 0.83. While this strong relationship suggests potential for developing basin-specific models and supports the integration of CM SAF products in further climate analyses, several factors must be considered. Extremely low values are observed for ET_{WB}^{GIRAFE}, which could bias the relationship and $ET_{CM\ SAF}$ shows a cluster of points near zero.

This cluster of low values is not unexpected, as it primarily corresponds to winter months when precipitation, evapotranspiration, and runoff are all typically reduced. As a result, the overall water balance, defined as the net difference between inputs and outputs, tends to be small. Under such conditions, even moderate uncertainties in individual components can strongly affect the residuals. For example, errors of approximately 10 mm in storage change estimates (Landerer and Swenson., 2012) and ~10% in discharge estimates (~5 mm month^{-1}; Aerts et al., 2024) can introduce significant relative variability. Additionally, ET from CM SAF tends to provide lower values (see section 2.2.1), which may contribute to the clustering near zero. This pattern is not limited to a single case but is also observed across most basins and for several other ET products. Finally, precipitation may be underestimated in the GIRAFE dataset. In the

Deleted:
Field Code Changed
Deleted: 2.7
Deleted: 6
Deleted: . However
Deleted: 01
Deleted: 4
Formatted: Subscript
Deleted: incorporating
Formatted: Font color: Text 1
Deleted: the number of valid matchups is 2.7 times lower while keeping similar
Deleted: the
Deleted: and uRMSD
Deleted: 2
Formatted: Font: (Default) Arial
Deleted: This strong relationship highlights the potential for developing basin-specific models, thereby enabling further climate analysis through the integration of CM SAF products. ¶ However, e
Formatted: English (US)
Deleted: .
Formatted: Font: (Default) Arial, 11 pt
Formatted: Font: (Default) Arial, 11 pt, Not Bold
Formatted: Font: (Default) Arial, 11 pt, Not Bold, Subscript
Formatted: Font: (Default) Arial, 11 pt
Formatted: Font: (Default) Arial
Formatted: Font: (Default) Arial
Formatted: Font: (Default) Arial, 11 pt
Formatted: Space After: 0 pt
Formatted: Font: (Default) Arial, 11 pt
Formatted: Superscript

case of GIRAFE, this underestimation is partly due to limitations in the GIRAFE v1 precipitation dataset, which has a reduced ability to detect precipitation in mid-to-high latitudes during snow or ice conditions. This issue arises from limitations in the passive microwave (PMW) input data and can lead to an underestimation of precipitation in these areas. A dedicated quality flag is available to help users identify and exclude potentially affected grid cells.

To further investigate the robustness of the observed relationships, we extended the R^2 analysis across all basins using different ET datasets, both with and without applying the GIRAFE snow flag. This was done to evaluate whether the improved agreement seen when using GIRAFE precipitation is consistent across datasets and basins, and to assess the potential role of snow-related biases.

Figure 10 displays the probability density functions (PDFs) of R^2 values between independent ET products (CM SAF, GLEAM, LSA SAF, ERA5-Land, GLDAS 2.1) and three water balance-based ET estimates: ET_{WB}^{GIRAFE}, ET_{WB}^{MSWEP}, and $ET_{WB}^{E\text{-}OBS}$. Across all datasets and basins, the use of GIRAFE precipitation systematically leads to higher R^2 values compared to MSWEP or E-OBS, indicating stronger alignment with independent ET products. This pattern is not limited to a specific product or region; it is observed across all independent ET datasets, suggesting that the higher correlations are not driven by dataset-specific features (e.g., CM SAF only) or by local conditions (e.g., the Düsseldorf basin). When applying the GIRAFE snow flag the number of valid months and basins decreases (see Figure 20 in Appendix), shifting the analysis toward summer periods. Nevertheless, the overall pattern remains: the use of GIRAFE precipitation continues to yield higher R^2 values, reinforcing the robustness of this result. This indicates that the observed improvements are not solely driven by extreme or low winter values but reflect a more consistent alignment between precipitation inputs and ET estimates.

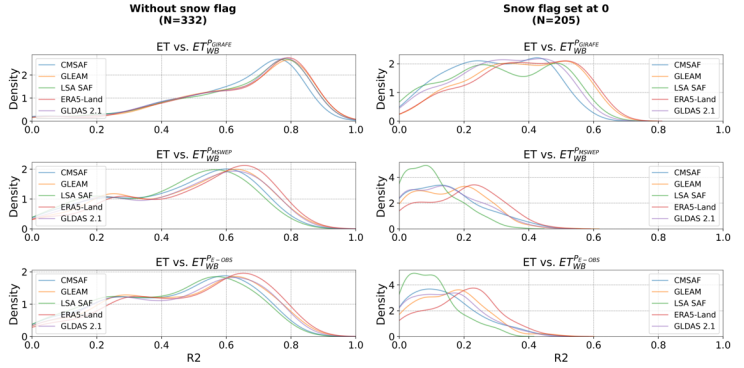


Figure 10: Probability density functions (PDFs) of R^2 values between independent evapotranspiration (ET) products (CM SAF, GLEAM, LSA SAF, ERA5-Land, GLDAS 2.1) and ET derived from the water balance-derived evapotranspiration estimates based using three different precipitation datasets: GIRAFE (top row, ET_{WB}^{GIRAFE}), MSWEP (middle row, ET_{WB}^{MSWEP}), and E-OBS (bottom row, $ET_{WB}^{E\text{-}OBS}$).Left panels show results without applying a snow flag, while right panels include only values where the GIRAFE snow flag is set to 0. Each color represents one ET product, with *N* the number of basins considered.

Formatted: Font: (Default) Arial, 11 pt

Deleted: As mentioned in Section 2.3,

Field Code Changed

Formatted: Font: 11 pt, Not Bold

Formatted: Superscript

Formatted: Normal, Left, Space Before: Auto, Line spacing: single

Formatted: Font: (Default) Arial, 11 pt

Formatted: Font: (Default) Arial, 11 pt

Formatted: Font: (Default) Arial, 11 pt

Formatted: Font: (Default) Arial, 11 pt

Formatted: Font: (Default) Arial, 11 pt

Formatted: Font: Arial

Formatted: Font: (Default) Arial, 11 pt

Formatted: Font: Arial

Formatted: Font: (Default) Arial, 11 pt

Formatted: Font: (Default) Arial, 11 pt

Formatted: Font: (Default) Arial, 11 pt, Not Bold

Formatted: Font: (Default) Arial, 11 pt

Formatted: Font: (Default) Arial, 11 pt

Formatted: Font: (Default) Arial, 11 pt

Field Code Changed

Formatted: Font: 11 pt, Not Bold

Formatted: Font: (Default) Arial, 11 pt

One possible hypothesis for this consistent increase in R^2 is that GIRAFE precipitation and the ET products share a dependence on satellite-based observations, which involve retrievals through the atmosphere and are influenced by large-scale atmospheric conditions (e.g., radiation, cloud cover, convective activity), which can introduce similar patterns in their outputs regardless of true variability. This can lead to higher R^2 values due to consistent seasonal signals in both precipitation and ET estimates. Another complementary hypothesis is that the higher R^2 values are driven by the stronger seasonal amplitude observed in the GIRAFE precipitation dataset. Among the three precipitation products, GIRAFE exhibits a markedly more pronounced seasonal cycle (see Figure 21 in Appendix), which more closely matches the seasonal variability in ET products. This stronger seasonal cycle in GIRAFE might reflect real climate patterns, but it could also be exaggerated, for example due to underestimation of winter precipitation. In either case, the stronger seasonal cycle aligns more closely with the dominant mode of variability in ET, which is also strongly seasonal, thereby inflating R^2 values. This interpretation is supported by a comparison of the annual cycle R^2 between precipitation and $ET_{CM\ SAF}$: values reach 0.77 for GIRAFE, but only 0.23 and 0.30 for E-OBS and MSWEP, respectively. A similar pattern is observed across all basins and ET products (results not shown), indicating that the stronger seasonal alignment between GIRAFE and ET products is not limited to a specific region. However, this hypothesis does not imply improved agreement in anomalies or short-term variability. Indeed, when comparing anomalies between ET_{WB} based on GIRAFE and $ET_{CM\ SAF}$, correlations are close to zero (not shown here). This suggests that the higher R^2 values may primarily reflect climatological alignment, such as annual cycles, rather than a more accurate representation of interannual or sub-seasonal dynamics.

This consistent improvement in correlation across datasets supports the potential use the ET_{WB}^{GIRAFE} vs. $ET_{CM\ SAF}$ relationship in practical applications, such as over the Düsseldorf basin. Indeed, this relationship could serve, for instance, as a basis for a gap-filling method or for extending time series of another component of the water balance equation, as illustrated in Figure 11. This figure shows the temporal course of $\frac{dS}{dt}$ over the Düsseldorf basin, derived from both GRACE solutions (CSR, JPL, GFZ and their ensemble mean) and computed from the water balance equation (Eq. 1) using the GIRAFE precipitation, GRDC discharge and evapotranspiration estimated from established linear relationship between ET_{WB}^{GIRAFE} and $ET_{CM\ SAF}$ ($ET_{WB}^{GIRAFE} = 1.2 \times ET_{CM\ SAF} - 42$). The water balance-derived $\frac{dS}{dt}$ captures the temporal dynamics well and aligns closely with GRACE estimations, especially during periods with data gaps, suggesting its usefulness for gap-filling. It seems important to note that there is no significant bias, which would not be the case if CM SAF evapotranspiration were used directly. While promising, this approach remains a simplified representation of storage dynamics and requires further refinement to improve its robustness and applicability. Despite its encouraging result, this approach is a simplified representation of $\frac{dS}{dt}$ and should be applied with caution. Indeed, it needs to be applied in a well-defined context in which the user is fully aware of the associated uncertainties and limitations. Its reliability depends on various factors as the selected GRACE solution used, the precipitation input to estimate ET_{WB}, and the characteristics of the study basin. The empirical relationship itself is subject to uncertainty and may not be applicable to all regions or climatic regimes. Note that the approach has also been tested after removing spurious clusters of points near zero, and the conclusions remain valid

Formatted: Font: (Default) Arial, 11 pt
Formatted: Justified, Line spacing: Multiple 1,15 li, Pattern: Clear
Formatted: Font: Not Bold
Formatted: Font: (Default) Arial, 11 pt
Formatted: Font: Not Bold
Formatted: Font: (Default) Arial, 11 pt
Formatted: Font: (Default) Arial, 11 pt
Formatted: Font: (Default) Arial, 11 pt, Not Bold
Formatted: Font: (Default) Arial, 11 pt
Field Code Changed
Formatted: Font: Not Bold
Formatted: Font: (Default) Arial, 11 pt
Formatted: Font: (Default) Arial, 11 pt
Formatted: Subscript
Formatted: Font: (Default) Arial, 11 pt
Formatted: Font: (Default) Arial, 11 pt
Formatted: Font: (Default) Arial, 11 pt
Formatted: Font: (Default) Arial, 11 pt, Subscript
Formatted: Font: (Default) Arial, 11 pt
Formatted: Subscript
Formatted: Font: (Default) Arial, 11 pt
Formatted: Font: (Default) Times New Roman, 12 pt, English (US)
Deleted: a limitation of the GIRAFE v1 precipitation dataset is its reduced ability to detect precipitation accurately in mid-to-high latitudes during snow or ice conditions, due to limitations in the passive microwave (PMW) input data. This can lead to the underestimation of precipitation in these areas. A dedicated quality flag is provided to help users identify and exclude potentially affected grid cells. Figure 10 illustrates the impact on the relation of applying different snow-day flags as criteria to extract the data. By limiting the number of snow days to 5 ... [2]
Formatted: Line spacing: Multiple 1,15 li
Formatted: Font: (Default) Arial, 11 pt
Formatted: Justified, Line spacing: Multiple 1,15 li
Formatted: Subscript
Formatted: Font: (Default) Arial, 11 pt
Formatted: Line spacing: Multiple 1,15 li
Deleted: ¶
Deleted: T
Deleted: be
Deleted: used
Deleted: the
Deleted: Figure 11
Deleted: 1
Deleted: 38
Deleted: with a snow-day flag threshold of maximum five ... [3]
Formatted: Font: (Default) Arial, 11 pt

(results not shown here). However, the users should define the acceptable error threshold for their specific application and check whether the error associated with the use of this gap-filling method falls within those limits.

In the following sections, we extend the comparison of ET estimates from various independent products with both ET_{WB} and ET_{PQ}, analyzing the results according to seasonal variability, basin size, and climatic regimes of the basin.

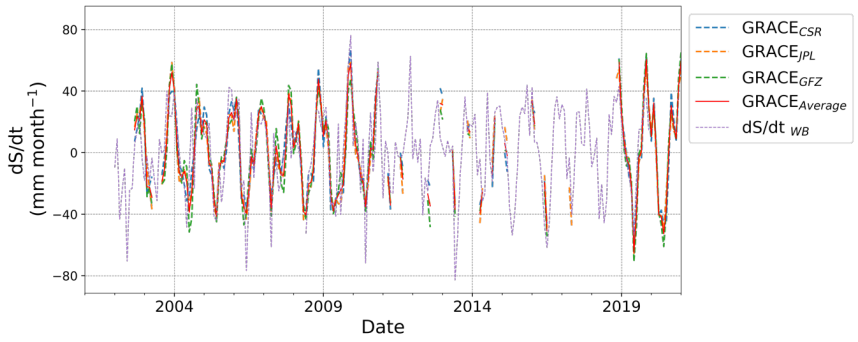


Figure 11: Temporal evolution of monthly $\frac{dS}{dt}$ (mm month^{-1}) estimates over the Düsseldorf basin, based on GRACE satellite observations and the water balance method. GRACE-derived $\frac{dS}{dt}$ includes individual solutions from CSR (blue dashed line), JPL (orange dashed line), and GFZ (green dashed line), along with their ensemble mean (solid red line). The water balance-based estimate ($\frac{dS}{dt}_{WB}$ dashed violet line), is calculated as P − ET_{WB} − Q, where ET_{WB} is derived from CM SAF ET data using the empirical relationship ET_{WB} = 1.2 × $ET_{CM\ SAF}$ − 42.

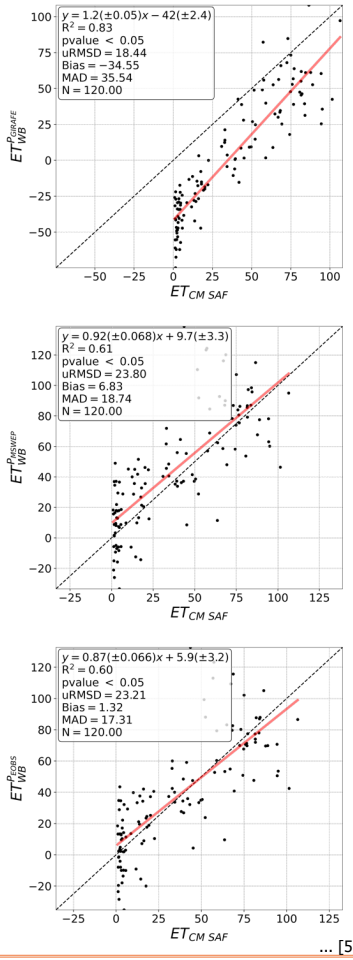
3.2 Monthly annual cycle analyses for the largest river stations

As observed in the previous section, strong seasonality is observed for the various datasets considered in the study. This seasonal sensitivity highlights the importance of conducting time-dependent analyses. Figure 12 displays the monthly mean annual cycle of ET derived from three different estimation approaches for the six largest stations of each river in our dataset with basin areas exceeding 90 000 km^2. For each basin, the blue, orange, and green solid lines represent the mean ET values derived from independent products (CM SAF, LSA SAF, GLEAM, ERA5-Land, GLDAS 2.1), from the water balance method with and without considering $\frac{dS}{dt}$ based on the three precipitation products, respectively. Shaded areas indicate the corresponding standard deviations showing the intra-product variability. Note that the standard deviation values for ET_{WB} and ET_{PQ} are identical, as they only depend on the choice of precipitation input used in their computation.

Across all basins, the highest intra-product variability, reflected by the standard deviation, is observed in summer for ET derived from independent products with average values around

Deleted: Therefore

Deleted: ¶
¶
¶
¶
¶
¶
¶
... [4]



Deleted: ... [5]

Formatted: Justified

Formatted: Caption

Deleted: −

Deleted: 38

Deleted: (see Figure 10).¶

Deleted: seven largest

15.3 ± 3 mm month^{-1}. At the opposite, variability is lowest for ET_{WB} and ET_{PQ} during the same period. Most of the variability is observed in winter and autumn, with standard deviation values around 18 mm month^{-1} when differences between values from GIRAFE and values from other products are higher (results not shown here). Overall, intra-product variability is higher for water balance-derived ET, with an average standard deviation of 14.7 mm month^{-1}, compared to 9.2 mm month^{-1} for ET from independent sources. This highlights the strong influence of the selected precipitation product on estimations. In addition, the magnitude of intra-product variability for ET_{WB} and ET_{PQ} is more dependent on the basin than for independent sources. For instance, the standard deviation in summer is only 6.1 mm month^{-1} at the Senta station, compared to 16.3 mm month^{-1} at the Lobith station. With regard to the product differences, as noted in the previous section, it appears that accounting for changes in water storage is crucial for temporal course analysis. The ET values from independent products are, for most of the cases, higher than those from ET_{WB} (on average 1.4 times higher), with the magnitude of this difference varying seasonally and with the basin. However, the general temporal pattern remains similar. In contrast, ET_{PQ} provides higher values in winter and autumn with values on average 2.7 and 1.04 times higher, respectively. In summer and spring, it shows much lower values, being 1.5 and 2.2 times lower on average. Consequently, the distinct bell-shaped curve observed for ET and ET_{WB} is no longer clearly apparent. The water storage change term plays a critical role, especially in summer, where its contribution is more important and can greatly impact the water balance closure.

Deleted: 4
Deleted: 7
Deleted: 3
Deleted: 3
Deleted: 8.9
Deleted: 3.7
Deleted: Novosaratovka
Deleted: 8
Deleted: 1

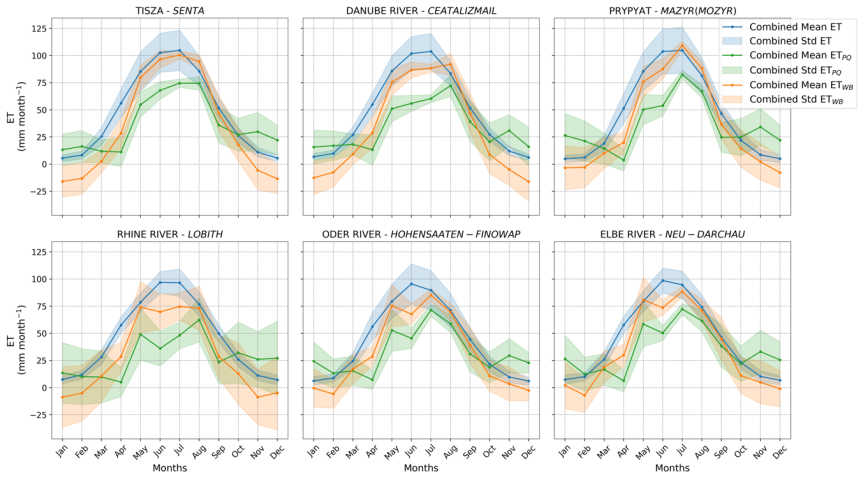


Figure 12: Monthly mean annual cycle (MMAC) of evapotranspiration (ET; mm month^{-1}) estimated from different approaches for the six largest river basins from our database. Each panel represents a different basin, illustrating the average (solid line) and the standard deviation (shaded area) of the evapotranspiration MMAC. The estimates are derived from independent datasets (in blue) and from the water balance equation using different precipitation products, with (orange) and without (green) considering changes in water storage.

Deleted: seven

Figure 13 and Figure 14 present the monthly variability of the imbalance (ϵ, Eq. 2) or excluding (Figure 14) the change in water storage ($\frac{dS}{dt}$). The imbalance is computed for each of the independent ET datasets: CM SAF, LSA SAF, GLEAM, ERA5-Land, and GLDAS 2.1. The analysis focuses on the six largest river stations in the database, with each subplot representing a different basin and illustrating the seasonal cycle of the imbalance. Overall, negative imbalances dominate across all basins for both ET_{WB}^{GIRAFE} and ET_{PQ}^{GIRAFE}, with only a few exceptions in certain months. On average, the imbalance is approximately -28 mm month^{-1}, with average values ranging from -15 mm month^{-1} for CM SAF to -35 mm month^{-1} for GLDAS 2.1 datasets. Thus, while the average imbalance may remain stable, the intra-annual variability and monthly magnitude of the imbalance vary across basins, both with and without accounting for changes in water storage. Those results reflect the combined influence of the temporal scale of analysis, input data uncertainty, potential errors in water balance components, discrepancies among independent ET products, and the effects of basin area and climatic regime.

Deleted: seven largest

Deleted: 7

Deleted: 4

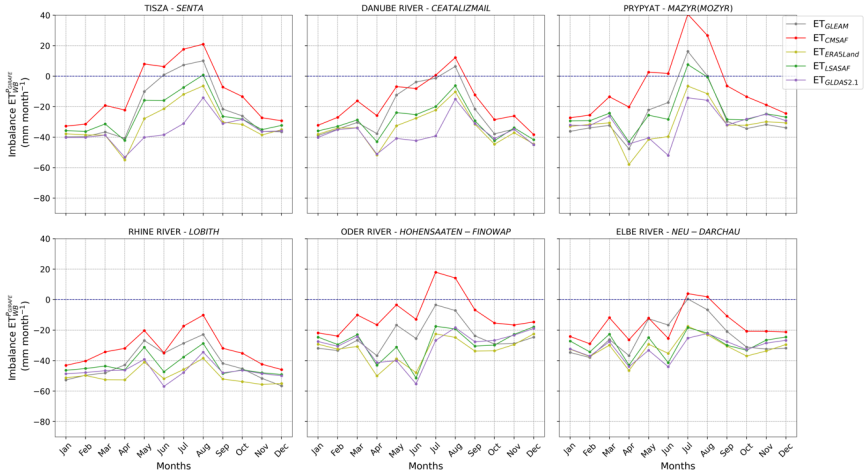


Figure 13: Monthly mean annual cycle of the imbalance (ϵ), calculated using Eq. 2, where each curve corresponds to a different independent ET dataset (CM SAF in red, LSA SAF in green, GLEAM in grey, ERA5-Land in yellow, and GLDAS 2.1 in violet). The GIRAFE dataset is used as input for precipitation. Results are shown for the six largest river basins in the dataset, with each panel representing a different basin. The dashed blue line indicates the zero-imbalance reference.

Deleted: seven largest

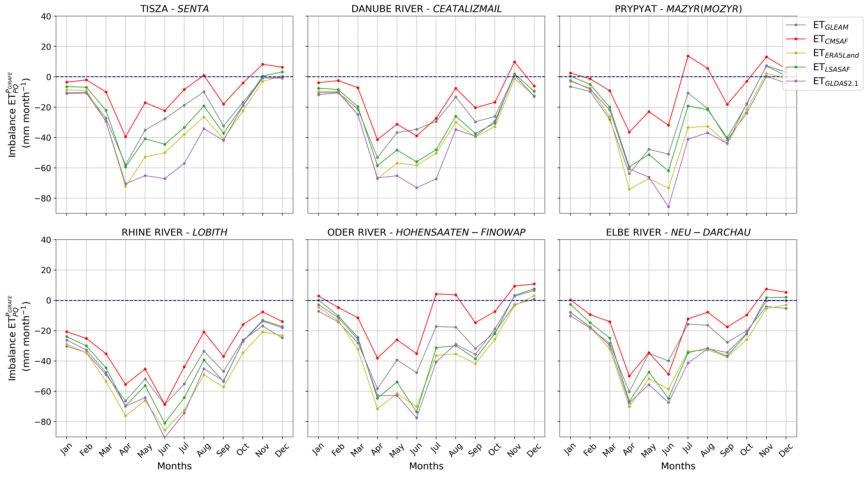


Figure 14: Same as Figure 13, but calculated without including the change in water storage ($\frac{dS}{dt}$) in Eq. 2.

3.3 All basins analysis

In the following section, this analysis is extended to the full dataset by considering not only the largest river basins, but also examining the influence of precipitation sources, basin size, and climate regime on the observed imbalances.

3.3.1 Imbalance

The averaged imbalance was calculated for each of the 332 basin time series using Eq. 2, based on different precipitation products (E-OBS, GIRAFE, MSWEP) and various ET sources (CM SAF, LSA SAF, GLEAM, ERA5-Land, and GLDAS 2.1). Figure 15 shows the distribution of the monthly mean imbalance (ϵ; Eq. 2) across each station in the database, comparing the impact of the precipitation product for the five independent ET products. Results are shown for the three precipitation datasets with or without the consideration of water storage changes. Overall, results tend to show negative imbalances, except while comparing with $ET_{CM\ SAF}$. The choice of precipitation dataset appears to have the largest impact on the mean imbalance, while the inclusion of the water storage term ($\frac{dS}{dt}$) has essentially no impact on the monthly mean imbalance when aggregated over the full dataset. This suggests that at this temporal

Deleted: 342

scale (mean across multiple years for each basin), changes in water storage tend not to be significant in analyzing the imbalance. Differences due to the inclusion or exclusion of $\frac{dS}{dt}$ are thus of secondary importance in this context. This observation aligns with findings in the literature (e.g., Zhang et al. (2023a)), but we refine this assumption by focusing specifically on the global imbalance from each basin and not the intra-annual variability discussed in the previous section. While assuming negligible storage changes may be valid in certain contexts, this assumption remains elusive and appears to be limited to well-defined study cases.

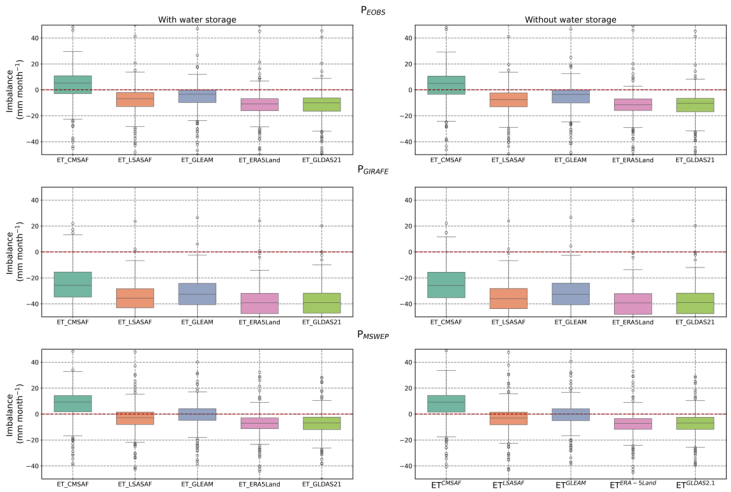


Figure 15: Boxplots of the monthly imbalance (ε; Eq. 2) averaged over the full time series for each of the 332 basins (N = 332). The imbalance is computed using three different precipitation datasets: E-OBS (top panel), GIRAFE (middle panel) and MSWEP (bottom panel). Evapotranspiration estimates are taken from CM SAF, LSA SAF, GLEAM, ERA5-Land, and GLDAS 2.1. The left panels show results including the water storage change component ($\frac{dS}{dt}$), while the right panels exclude it. The dashed red line indicates the zero-imbalance reference.

Deleted: 342

Deleted: 342

The second-order effect comes from the variability between independent ET products, which nevertheless significantly influence the imbalance values. Among these, results for CM SAF ET dataset tend to provide higher imbalance values than other sources with median values, on average 12.5 mm month^{-1} higher. Such results were expected as they reflect the lower values observed by Moutier et al. (2023a) over Europe while making inter-comparisons between $ET_{CM\ SAF}$ and other products. Importantly, the choice of the precipitation dataset strongly determines the magnitude of the imbalance. MSWEP consistently produces the

Deleted: 7

smallest average imbalances across most of the ET products and configurations. On average median values are of -34.4 mm month^{-1} for GIRAFE, -5.3 mm month^{-1} for E-OBS, and -1.6 mm month^{-1} for MSWEP. This suggests that MSWEP may offer the most consistent agreement in a water balance framework at this scale. However, the best-performing precipitation dataset still depends on the independent ET source considered, indicating no universally optimal combination. For instance, when using the $ET_{CM\ SAF}$ dataset, the best performances for ET_{WB} are obtained while using E-OBS as precipitation dataset to get a median value of 5.3 mm month^{-1} when it is of 9.2 mm month^{-1} while using MSWEP as precipitation dataset.

Deleted: 3.
Deleted: 7
Deleted: 5
Deleted: 5
Deleted: 3

While the precipitation dataset appears to dominate the behaviour of the imbalance at the global scale, these aggregated results may mask important patterns. Therefore, we investigated how the imbalance varies according to basin size and climate regime, using various precipitation datasets.

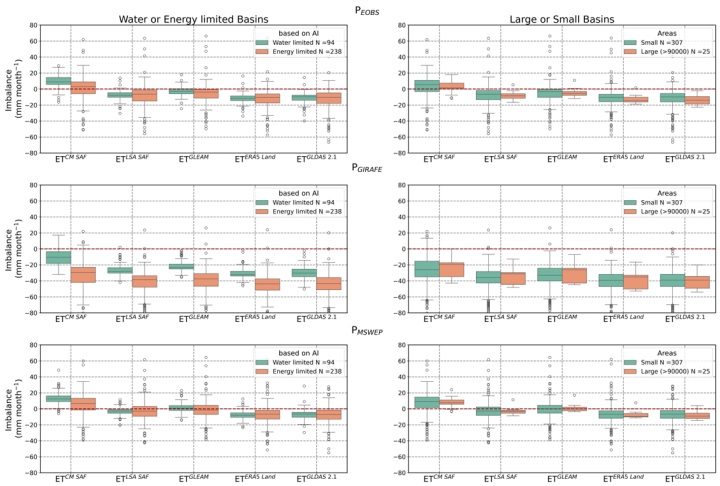


Figure 16: Boxplots of the monthly imbalance (ϵ; Eq. 2) averaged over the full time series of each basin using Eq. 2, based on different precipitation products: E-OBS (top panel), GIRAFE (middle panel) and MSWEP (bottom panel). Evapotranspiration estimates are taken from CM SAF, LSA SAF, GLEAM, ERA5-Land, and GLDAS 2.1. Each panel contains two boxplots per dataset, distinguishing between different basin characteristics. Left panels categorize basins based on the aridity index (AI), with energy-limited (wet) basins and water-limited (dry) basins. The bottom panels differentiate between large basins (≥ 90 000 km^2) and small basins (< 90 000 km^2). The dashed red line indicates the zero-imbalance reference.

Figure 16 presents the distribution of monthly imbalance values computed using three precipitation products (E-OBS, GIRAFE, MSWEP), across different subsets: water-limited

versus energy-limited basins and large versus small basins, for all independent ET datasets (CM SAF, LSA SAF, GLEAM, ERA5-Land, GLDAS 2.1). As previously observed, the choice of the precipitation product remains the primary driver of the imbalance magnitude and imbalance values for $ET_{CM\ SAF}$ tend to be higher. Nevertheless, basin characteristics also introduce significant differences. The most important distinction is observed between water-limited and energy-limited basins while using the GIRAFE dataset. The difference in median imbalance for water-limited and energy-limited basins is between 10 and 19 mm month^{-1} for the LSA SAF and the CM SAF datasets, respectively. For E-OBS and MSWEP, a smaller difference around 6 and 5.5 mm month^{-1} are observed for $ET_{CM\ SAF}$, while other ET datasets show median differences within ± 2 mm month^{-1}. These results suggest that $ET_{CM\ SAF}$ estimations tend to be more sensitive to the climatic conditions of the basin as compared to other datasets. In contrast, the influence of basin size (large vs. small) appears to be of secondary importance for all configurations. Only the imbalance based on GIRAFE and ET from CM SAF and GLEAM are slightly impacted (≥ 5 mm month^{-1}), with median differences between large and small basins of 6.7 and 7 mm month^{-1}, respectively. These variations are smaller than those observed between ET products, highlighting that inter-product variability dominates. This limited sensitivity of the imbalance value according to the basin area is expected, as the primary goal of focusing on large basins is to ensure more reliable estimations of $\frac{dS}{dt}$. Moreover, this finding aligns with previous studies such as Zhang et al. (2012) and Zhang et al. (2023a), who reported no significant differences in results when separating basins into small (≤5000 km^2) and large (> 5000 km^2) subsets. However, as observed in Figure 15, taking into account the water storage term or not does not led to significant changes in the resulting imbalance values when aggregated at the basin scale. It is worth noting that it remains difficult to identify a single best-performing independent ET product based on basin size or climate regime. The relative performance of each product depends firstly on the precipitation dataset used and secondly on the climate regime for certain cases, indicating a complex interaction between input datasets and local conditions. Notably, when using the GIRAFE dataset and/or $ET_{CM\ SAF}$, the climate regime of the basin plays a crucial role in shaping the water balance closure performance. In contrast, basin size plays a minor role, and the effect of including $\frac{dS}{dt}$ remains limited at this aggregation scale.

Deleted: 8

Deleted: is

Deleted: ,

Deleted: , and LSA SAF

Formatted: Superscript

Deleted: 7.2

Deleted: ,

Deleted: , and 5

3.3.2 Statistical analysis

To further explore the influence of water storage changes without focusing the analysis on the imbalance, we present a statistical comparison between the monthly CM SAF evapotranspiration product ($ET_{CM\ SAF}$) and the water balance-derived estimates computed both with (ET_{WB}) and without (ET_{PQ}) the change in water storage component for the three precipitation datasets, using all available match-ups across all basins (Figure 17).

Specifically, with and without considering $\frac{dS}{dt}$, bias values are approximately 28 mm month^{-1} (GIRAFE), –2 mm month^{-1} (E-OBS), and –7 mm month^{-1} (MSWEP). Accounting for $\frac{dS}{dt}$ increases the coefficient of determination (R^2) from 0.18 to 0.43 for GIRAFE, from 0.07 to 0.34 for E-OBS, and from 0.08 to 0.36 for MSWEP. Thus, the highest agreement is observed with

Deleted: 7

Deleted: 3

Deleted: 4

the GIRAFE precipitation product. When limiting the analysis to the six largest river basins (where $\frac{dS}{dt}$ estimates are considered more reliable) the R^2 reached values of 0.76 with GIRAFE (see Figure 22 in the Appendix). This analysis highlights the added value of including water storage changes in the water balance, particularly in configurations where

Deleted: seven largest

Deleted: 7

Deleted: Figure 20

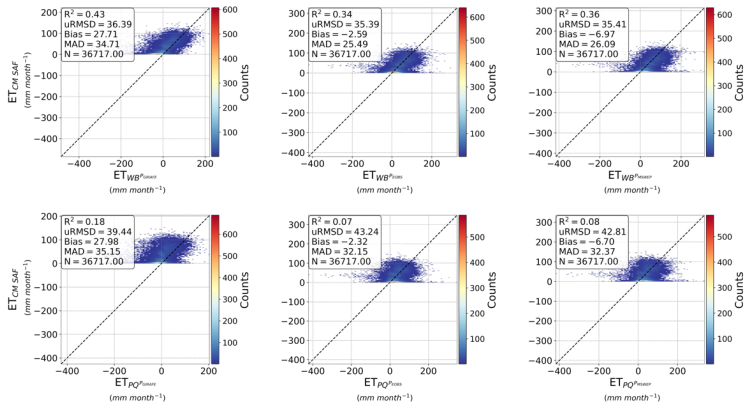


Figure 17: Density scatter plots comparing monthly evapotranspiration estimates from the CM SAF product ($ET_{CM\ SAF}$ in mm month^{-1}) with water balance-derived evapotranspiration estimates with (ET_{WB} in mm month^{-1}; top panels) and without (ET_{PQ} in mm month^{-1}; bottom panels) considering $\frac{dS}{dt}$ for all basins. GIRAFE (left panels), E-OBS (middle panels), and MSWEP (right panels) precipitation products are considered for the calculation of ET_{WB} and ET_{PQ}. The color scale indicates point density, with warmer colors representing higher densities. The dashed black line represents the 1:1 relationship. Statistical metrics, including the coefficient of determination (R^2), unbiased root mean square deviation (uRMSD), bias, mean absolute deviation (MAD), and the number of observations (N), are provided in each panel.

storage dynamics have a greater influence on evapotranspiration estimates. Moreover, the stronger performance observed for larger basins suggests that global statistics may mask patterns.

To investigate this further, similar statistical analysis has been applied, using the GIRAFE precipitation, to all other evapotranspiration sources (LSA SAF, GLEAM, ERA5-Land, and GLDAS 2.1) and across five different basin subsets: all basins, water-limited basins (AI > 1), energy-limited basins (AI ≤ 1), small basins (< 90 000 km^2), and large basins (≥ 90 000 km^2). Figure 18 summarizes the results using heatmaps for several performance metrics, including MAD, bias, R^2, uRMSD, and N. Across all basin subsets, CM SAF consistently outperforms other datasets in terms of bias and MAD. The best performances are observed in water-limited basins, where CM SAF shows a bias of 9 mm month^{-1} and MAD of 22 mm month^{-1}. For similar subset, other products exhibit bias between 21 and 30 mm month^{-1} and MAD values between

Formatted: Superscript

27 and 33 mm month^{-1}. It is worth noting that incorporating water storage changes ($\frac{dS}{dt}$) does not significantly reduce MAD or bias, suggesting that the main discrepancies between ET from the water balance and satellite/reanalysis products are not primarily due to the omission of the storage term. Instead, errors in precipitation (P), discharge (Q), or biases in the ET datasets themselves are likely responsible for most of the differences.

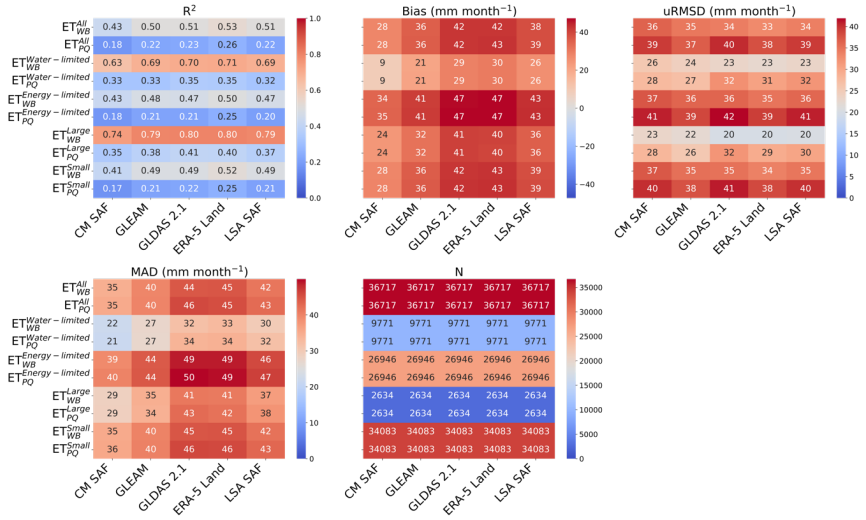


Figure 18: Heatmaps of statistical metrics assessing the agreement between monthly evapotranspiration estimates from different datasets (CM SAF, GLEAM, GLDAS 2.1, ERA5-Land, and LSA SAF) and water balance-derived evapotranspiration estimates considering (ET_{WB}) or not (ET_{PQ}) $\frac{dS}{dt}$ and based on the GIRAFE precipitation product. Statistics are made for all matchup-ups include in different subsets: All, water-limited (AI > 1), energy-limited (AI ≤ 1), small (< 90 000 km²) and large (≥ 90 000 km²). The panels display (top-left) the coefficient of determination (R), (top-middle) bias (mm month^{-1}), (top-right) unbiased root mean square deviation (uRMSD, mm month^{-1}), (bottom-left) median absolute deviation (MAD, mm month^{-1}), and (bottom-right) the number of observations (N). Warmer colors indicate higher values, while cooler colors represent lower values.

In terms of uRMSD, results vary between 20 mm month^{-1} for ET_{WB} in large basins and 42 mm month^{-1} for ET_{PQ} in energy-limited basins. uRMSD values are generally slightly better for ET_{WB} compared to ET_{PQ}, but marginal improvements of 5 mm month^{-1} can still be observed in specific cases. No independent ET product consistently outperforms others across all subsets for this metric.

Regarding R^2, values are generally in the same order of magnitude across independent datasets within each subset, but become significantly higher when water storage variations are included (ET_{WB}). The improvement in R^2 (especially for large basins) shows that adding $\frac{dS}{dt}$

improves the ability to capture the temporal variability of ET. This reinforces the idea that storage is an important component to analyze inter-annual and seasonal fluctuations, even if its effect is not significant on cumulative ET values (no impact on the bias). This improvement of R^2 is most notable in large basins (0.74 for ET_{WB}^{Large} and 0.35 for ET_{PQ}^{Large}), although small basins also benefit, but with less improvement and lower absolute values (0.41 for ET_{WB}^{Small} and 0.17 for ET_{PQ}^{Small}). This observation for small basins is potentially due to their faster hydrological response times and less significant storage capacity, meaning that short-term precipitation and discharge dominate the water balance. Finally, water-limited basins tend to produce higher R^2 values than energy-limited ones. For instance, in the case of the CM SAF dataset, R^2 are 0.43 and 0.63 for energy-limited and water-limited basins, respectively. Additionally, the improvement of the R^2 is more pronounced for water-limited basins than energy-limited ones. This is potentially due to the fact that the influence of $\frac{dS}{dt}$ on the evapotranspiration in energy-limited basins is less direct because ET is primarily controlled by energy availability rather than water storage. Indeed, the temporal course in energy-limited basins is mainly driven by the availability in energy. Overall, the R^2 values obtained in this study are of the same order of magnitude as those reported in previous works such as (Ruhoff et al., 2022; Zhang et al., 2023a). In addition, Ruhoff et al. (2022) also observed better consistency in water-limited conditions. However, contrasting results were found by Zhang et al. (2023a), who observed higher R^2 values in energy-limited basins. This discrepancy can be attributed to several factors. First, each basin has its own unique hydrological and climatic characteristics that may affect the estimations performance. Second, differences in the input datasets used in the water balance equation (precipitation, discharge, or water storage) which can lead to different results. Third, a key limitation in classifying basins as water- or energy-limited lies in the use of the aridity index, which is sensitive to both its mathematical formulation and the choice of input datasets (e.g., reference evapotranspiration and precipitation sources). As a result, a basin classified as water-limited in one study could be considered energy-limited in another, potentially leading to differing interpretations of the same region's water balance behaviour.

It is also important to note that calculating R^2 values for every basin and then averaging them for each subset results in higher aggregated R^2 values (see Figure 23 in the Appendix). For example, in the large basin subset and when considering water storage change, the mean R^2 reaches approximately 0.82 (0.74 in Figure 18). In addition, the difference in R^2 values between water-limited and energy-limited basins diminishes to about 0.06 (0.67 vs. 0.61, respectively) when compared against $ET_{CM\ SAF}$.

Deleted: 2
Deleted: 8
Deleted: 4
Deleted: water-limited and
Deleted: Figure 21
Deleted: 4
Deleted: 7
Deleted: 8

4 Conclusion

This study presents a comprehensive evaluation of the CM SAF LANDFLUX Ed. 1 evapotranspiration (ET) product within a water balance framework, using data from 332 European river basins. One of the key objectives is to assess the potential of the CM SAF ET product for hydrological basin applications, particularly in the context of water balance closure. Several ET datasets (CM SAF, LSA SAF, ERA5-Land, GLDAS 2.1, and GLEAM) and three precipitation products (GIRAFE, E-OBS, and MSWEP) were assessed to analyze water balance closure performance, quantify the residual (imbalance), and explore the sensitivity of estimates to basin characteristics and input choices.

A major focus of this study was the role of water storage changes ($\frac{dS}{dt}$), derived from GRACE data, in the water balance equation. While the inclusion of $\frac{dS}{dt}$ had a limited effect on the long-term average imbalance across all basins and the full study period, it is crucial to take it into account for accurately capturing intra-annual variations and catching the temporal course of independent evapotranspiration products such as CM SAF. For instance, when considering all available match-ups, the correlation (R^2) between CM SAF ET and water-balance-derived ET was relatively low without accounting for $\frac{dS}{dt}$ (e.g., R^2 = 0.18 with GIRAFE). However, when $\frac{dS}{dt}$ was explicitly accounted for, more robust relationship was obtained (R^2 = 0.43 overall and up to R^2 = 0.76 for largest river basins). This highlights that assuming $\frac{dS}{dt}$ to be negligible over the time period can lead to weaker temporal consistency, even when long-term mean imbalances remain unaffected.

More broadly, the choice of precipitation dataset was the primary driver of imbalance, reaffirming the first-order importance of accurate precipitation estimates in closing the water balance and underlining the challenges in estimating precipitation, especially regarding snow cases for GIRAFE. However, the best-performing configuration depends on the combination of ET and precipitation products. For example, GIRAFE showed the strongest consistency (i.e., higher R^2) with independent ET products, whereas MSWEP tended to minimize the average imbalance across most ET products. These findings indicate that no universally optimal combination exists. Therefore, the selection of input products should be guided by the specific objectives or/and the users' needs. Furthermore, the variations of the results depending on the chosen input indicate that independent ET datasets, such as the CM SAF ET product, may offer greater reliability than water-balance derived ET estimates.

The influence of basin size on imbalance was limited, especially when compared to the climatic regime. Differences in performance were more pronounced between energy-limited and water-limited basins. For example, when using GIRAFE precipitation and CM SAF ET, the median imbalance differed by up to 19 mm month^{-1} between the two climate regimes. However, the classification of basins by aridity index (AI) itself is sensitive to the choice of input datasets (e.g., reference evapotranspiration and precipitation sources), and thus this climate-based distinction must be interpreted with caution.

Overall, the CM SAF LANDFLUX Ed. 1 product demonstrated strong potential for hydrological applications, especially when combined with the GIRAFE precipitation dataset. While MSWEP emerged as a suitable compromise for minimizing long-term water balance imbalances, the

Deleted: 342

Deleted: taken into account,

Deleted: 4

Deleted: 82

Deleted: This highlights that omitting the change in water storage leads to weaker temporal consistency, even when long-term mean imbalances remain unaffected.

Deleted: 8

combination of CM SAF ET and GIRAFE precipitation proved to be the best-performing pair for capturing temporal dynamics, showing both high correlation and low mean absolute differences. Building on this, we have shown the potential of a gap-filling approach using an empirical linear relationship between CM SAF ET and water-balance ET derived from GIRAFE data. This approach captures the temporal variability of $\frac{dS}{dt}$ well, particularly for filling GRACE data gaps. The analysis was conducted for a typical example at the Düsseldorf station in the Rhine River. However, it must be applied carefully, considering the basin characteristics, precipitation accuracy, and quality of available data. Users should always define their acceptable uncertainty thresholds before applying such empirical corrections.

Additionally, the CM SAF LANDFLUX Ed. 1 evapotranspiration product, which combines high temporal and spatial resolution over nearly 40 years, has proven to be a reliable, well performing, and accessible tool, making it highly suitable for large-scale hydrological and climatological studies. The results highlight its potential for both scientific research and operational water resource management. The study also illustrates the value of integrating satellite-based datasets and encourages the continued refinement of the methodologies for gap-filling, uncertainty quantification, and application-specific optimization.

5 Acknowledgements

The authors would like to thank Nicolas Ghilain, Hannes Konrad and Almudena Velazquez for their support and suggestions in writing this manuscript.

6 References

Adler, R. F., Huffman, G. J., Chang, A., Ferraro, R., Xie, P.-P., Janowiak, J., Rudolf, B., Schneider, U., Curtis, S., Bolvin, D., et al. (2003). The version-2 global precipitation climatology project (gpcp) monthly precipitation analysis (1979–present). Journal of hydrometeorology, 4(6):1147–1167.

Aerts, J. P., Hoch, J. M., Coxon, G., Van De Giesen, N. C., & Hut, R. W. (2024). On the importance of discharge observation uncertainty when interpreting hydrological model performance. *Hydrology and Earth System Sciences*, *28*(22), 5011-5030.

Balsamo, G., Beljaars, A., Scipal, K., Viterbo, P., van den Hurk, B., Hirschi, M., and Betts, A. K. (2009). A revised hydrology for the ecmwf model: Verification from field site to terrestrial water storage and impact in the integrated forecast system. Journal of hydrometeorology, 10(3):623–643.

Barrios, J., Arboleda, A., Dutra, E., Trigo, I., and Gellens-Meulenberghs, F. (2024). Evapotranspiration and surface energy fluxes across europe, africa and eastern south america throughout the operational life of the meteosat second generation satellite. Geoscience Data Journal, 11(4):589–607.

Beck, H. E., Van Dijk, A. I., Levizzani, V., Schellekens, J., Miralles, D. G., Martens, B., and De Roo, A. (2017). Mswep: 3-hourly 0.25 global gridded precipitation (1979–2015) by merging gauge, satellite, and reanalysis data. Hydrology and Earth System Sciences, 21(1):589–615.

Beck, H. E., Wood, E. F., Pan, M., Fisher, C. K., Miralles, D. G., Van Dijk, A. I., McVicar, T. R., and Adler, R. F. (2019). Mswep v2 global 3-hourly 0.1 precipitation: methodology and quantitative assessment. Bulletin of the American Meteorological Society, 100(3):473–500.

Behrendt, A., Wulfmeyer, V., Senff, C., Muppa, S. K., Späth, F., Lange, D., Kalthoff, N., and Wieser, A. (2020). Observation of sensible and latent heat flux profiles with lidar. Atmospheric Measurement Techniques, 13(6):3221–3233.

Best, M. J., Abramowitz, G., Johnson, H., Pitman, A., Balsamo, G., Boone, A., Cuntz, M., Decharme, B., Dirmeyer, P., Dong, J., et al. (2015). The plumbing of land surface models: benchmarking model performance. Journal of Hydrometeorology, 16(3):1425–1442.

Bosshard, T., Kotlarski, S., Zappa, M., and Schär, C. (2014). Hydrological climate-impact projections for the rhine river: Gcm–rcm uncertainty and separate temperature and precipitation effects. Journal of Hydrometeorology, 15(2):697–713.

Chambon, P., Jobard, I., Roca, R., and Viltard, N. (2013). An investigation of the error budget of tropical rainfall accumulation derived from merged passive microwave and infrared satellite measurements. Quarterly Journal of the Royal Meteorological Society, 139(673):879893.

Chen, J., Tapley, B., Rodell, M., Seo, K.-W., Wilson, C., Scanlon, B. R., and Pokhrel, Y. (2020). Basin-scale river runoff estimation from grace gravity satellites, climate models, and in situ observations: A case study in the amazon basin. Water resources research, 56(10):e2020WR028032.

Formatted: Justified

Formatted: Font: Times New Roman, 12 pt

Formatted: Left, Space After: 0 pt, Line spacing: single

Cook, B. I., Smerdon, J. E., Cook, E. R., Williams, A. P., Anchukaitis, K. J., Mankin, J. S., Allen, K., Andreu-Hayles, L., Ault, T. R., Belmecheri, S., et al. (2022). Megadroughts in the common era and the anthropocene. Nature Reviews Earth & Environment, 3(11):741–757.

Cornes, R. C., Van Der Schrier, G., Van Den Besselaar, E. J., and Jones, P. D. (2018). An ensemble version of the e-obs temperature and precipitation data sets. Journal of Geophysical Research: Atmospheres, 123(17):9391–9409.

Derber, J. C., Parrish, D. F., and Lord, S. J. (1991). The new global operational analysis system at the national meteorological center. Weather and forecasting, 6(4):538–547.

Eyring, V., Gleckler, P. J., Heinze, C., Stouffer, R. J., Taylor, K. E., Balaji, V., Guilyardi, E., Joussaume, S., Kindermann, S., Lawrence, B. N., et al. (2016). Towards improved and more routine earth system model evaluation in cmip. Earth System Dynamics, 7(4):813–830.

Fisher, J. B., Melton, F., Middleton, E., Hain, C., Anderson, M., Allen, R., Mccabe, M. F., Hook, S., Baldocchi, D., Townsend, P. A., Kilic, A., Tu, K., Miralles, D., Perret, J., Lagouarde, J.-P., Waliser, D., Purdy, A. J., French, A., Schimel, D., Famiglietti, J. S., Stephens, G., and Wood, E. F. (2017). The future of evapotranspiration: Global requirements for ecosystem functioning, carbon and climate feedbacks, agricultural management, and water resources. Water Resources Research, 10.1002/20:2618–2626.

Ghilain, N., Arboleda, A., and Gellens-Meulenberghs, F. (2011). Evapotranspiration modelling at large scale using near-real time msg seviri derived data. Hydrology and Earth System Sciences, 15(3):771–786.

Ghilain, N., Arboleda, A., Sepulcre-Cantò, G., Batelaan, O., Ardö, J., and GellensMeulenberghs, F. (2012). Improving evapotranspiration in a land surface model using biophysical variables derived from msg/seviri satellite. Hydrology and Earth System Sciences, 16(8):2567–2583.

Görgen, K., Beersma, J., Brahmer, G., Buiteveld, H., Carambia, M., de Keizer, O., Krahe, P., Nilson, E., Lammersen, R., Perrin, C., et al. (2010). Assessment of climate change impacts on discharge in the rhine river bassin: results of the rheinblick2050 project. Technical report, Research Report] irstea.

Han, J., Yang, Y., Roderick, M. L., McVicar, T. R., Yang, D., Zhang, S., and Beck, H. E. (2020). Assessing the steady-state assumption in water balance calculation across global catchments. Water Resources Research, 56(7):e2020WR027392.

Hasenmueller, E. A. and Criss, R. E. (2013). Water balance estimates of evapotranspiration rates in areas with varying land use. Evapotranspiration—An Overview, pages 1–21.

Haylock, M., Hofstra, N., Klein Tank, A., Klok, E., Jones, P., and New, M. (2008). A european daily high-resolution gridded data set of surface temperature and precipitation for 19502006. Journal of Geophysical Research: Atmospheres, 113(D20).

Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Dee, D., Horányi, A., Nicolas, J., Peubey, C., Radu, R., Rozum, I., et al. (2019). The era5 global atmospheric reanalysis at ecmwf as a comprehensive dataset for climate data homogenization, climate variability, trends and extremes. In Geophysical Research Abstracts, volume 21.

Hobbins, M. T., Ramirez, J. A., and Brown, T. C. (2001). The complementary relationship in estimation of regional evapotranspiration: An enhanced advection-aridity model. Water Resources Research, 37(5):1389–1403.

Huffman, G. J., Adler, R. F., Morrissey, M. M., Bolvin, D. T., Curtis, S., Joyce, R., McGavock, B., and Susskind, J. (2001). Global precipitation at one-degree daily resolution from multisatellite observations. Journal of hydrometeorology, 2(1):36–50.

Joetzjer, E., Douville, H., Delire, C., Ciais, P., Decharme, B., and Tyteca, S. (2013). Hydrologic benchmarking of meteorological drought indices at interannual to climate change timescales: a case study over the amazon and mississippi river basins. Hydrology and Earth System Sciences, 17(12):4885–4895.

Kim, D. and Rhee, J. (2016). A drought index based on actual evapotranspiration from the bouchet hypothesis. Geophysical Research Letters, 43(19):10–277.

Klein Tank, A. M., Wijngaard, J., Können, G., Böhm, R., Demarée, G., Gocheva, A., Mileta, M., Pashiardis, S., Hejkrlik, L., Kern-Hansen, C., et al. (2002). Daily dataset of 20th-century surface air temperature and precipitation series for the european climate assessment. International Journal of Climatology: A Journal of the Royal Meteorological Society, 22(12):14411453.

Klok, E. and Klein Tank, A. (2009). Updated and extended european dataset of daily climate observations. International Journal of Climatology: A Journal of the Royal Meteorological Society, 29(8):1182–1191.

Konrad, H., Roca, R., Niedorf, A., Finkensieper, S., Schröder, M., Cloché, S., ... & Hollmann, R. (2025). GIRAFE v1: a global climate data record for precipitation accompanied by a daily sampling uncertainty. *Earth System Science Data*, *17*(8), 4097-4124.

Konrad, H., Schröder, M., Roca, R., and Niedorf, A. (2024). Validation report: Global interpolated rainfall estimation, version 1 (girafe v1). Technical Report SAF/CM/DWD/VAL/GIRAFE/1.1, EUMETSAT Satellite Application Facility on Climate Monitoring (CM SAF). DOI: 10.5676/EUM_SAF_CM/GIRAFE/V001.

Kukal, M. and Irmak, S. (2016). Long-term patterns of air temperatures, daily temperature range, precipitation, grass-reference evapotranspiration and aridity index in the usa great plains: Part ii. temporal trends. Journal of Hydrology, 542:978–1001.

Kumar, S. V., Peters-Lidard, C. D., Tian, Y., Houser, P. R., Geiger, J., Olden, S., Lighty, L., Eastman, J. L., Doty, B., Dirmeyer, P., et al. (2006). Land information system: An interoperable framework for high resolution land surface modeling. Environmental modelling & software, 21(10):1402–1415.

Formatted: Font: (Default) Arial, 11 pt

Formatted: Justified

Deleted: Konrad, H., Roca, R., Niedorf, A., Finkensieper, S., Schröder, M., Cloché, S., Panegrossi, G., Sanò, P., Kidd, C., Jucá Oliveira, R. A., et al. (2025). Girafe v1: a global climate data record for precipitation accompanied by a daily sampling uncertainty. Earth System Science Data Discussions, 2025:1–40.¶

Landerer, F. W., & Swenson, S. C. (2012). Accuracy of scaled GRACE terrestrial water storage estimates. *Water resources research*, *48*(4).

Landerer, F. (2021a). Tellus_grac_l3_csr_rl06_lnd_v04. Dataset accessed 2025-03-05.

Landerer, F. (2021b). Tellus_grac_l3_gfz_rl06_lnd_v04. Dataset accessed 2025-03-05.

Landerer, F. (2021c). Tellus_grac_l3_jpl_rl06_lnd_v04. Dataset accessed 2025-03-05.

Landerer, F. (2023a). Gfz tellus grace-fo level-3 monthly land water-equivalent-thickness surface mass anomaly release 6.3 version 04. Dataset accessed 2025-03-05.

Landerer, F. (2023b). JPL TELLUS GRACE-FO Level-3 Monthly Land Water-EquivalentThickness Surface Mass Anomaly Release 6.3 version 04. Dataset accessed 2025-03-05.

Landerer, F. (2024). Csr tellus grace-fo level-3 monthly land water-equivalent-thickness surface mass anomaly release 6.3 version 04. Dataset accessed 2025-03-05.

Landerer, F. W. and Swenson, S. (2012). Accuracy of scaled grace terrestrial water storage estimates. Water resources research, 48(4).

Lian, X., Piao, S., Huntingford, C., Li, Y., Zeng, Z., Wang, X., Ciais, P., McVicar, T. R., Peng, S., Ottlé, C., et al. (2018). Partitioning global land evapotranspiration using cmip5 models constrained by observations. Nature Climate Change, 8(7):640–646.

Liou, Y.-A. and Kar, S. K. (2014). Evapotranspiration estimation with remote sensing and various surface energy balance algorithms—a review. Energies, 7(5):2821–2849.

Liu, J., Zhang, J., Kong, D., Feng, X., Feng, S., and Xiao, M. (2021). Contributions of anthropogenic forcings to evapotranspiration changes over 1980–2020 using gleam and cmip6 simulations. Journal of Geophysical Research: Atmospheres, 126(22):e2021JD035367.

Liu, W., Wang, L., Zhou, J., Li, Y., Sun, F., Fu, G., Li, X., and Sang, Y.-F. (2016). A worldwide evaluation of basin-scale evapotranspiration estimates against the water balance method. Journal of Hydrology, 538:82–95.

Martens, B., Miralles, D. G., Lievens, H., Van Der Schalie, R., De Jeu, R. A., Fernández-Prieto, D., Beck, H. E., Dorigo, W. A., and Verhoest, N. E. (2017). Gleam v3: Satellite-based land evaporation and root-zone soil moisture. Geoscientific Model Development, 10(5):19031925.

McCabe, M. F., Ershadi, A., Jimenez, C., Miralles, D. G., Michel, D., and Wood, E. F. (2016). The gewex landflux project: Evaluation of model evaporation using tower-based and globally gridded forcing data. Geoscientific Model Development, 9(1):283–305.

Michailovsky, C. I., Coerver, B., Mul, M., and Jewitt, G. (2023). Investigating sources of variability in closing the terrestrial water balance with remote sensing. Hydrology and Earth System Sciences, 27(23):4335–4354.

Michel, D., Jiménez, C., Miralles, D. G., Jung, M., Hirschi, M., Ershadi, A., Martens, B., McCabe, M. F., Fisher, J. B., Mu, Q., et al. (2016). The wacmos-et project–part 1: Tower-scale

Formatted: Font: (Default) Arial, 11 pt
Formatted: Justified
Formatted: Font: Times New Roman, 12 pt
Formatted: Left, Space After: 0 pt, Line spacing: single

evaluation of four remote-sensing-based evapotranspiration algorithms. Hydrology and Earth System Sciences, 20(2):803–822.

Miralles, D. G., Bonte, O., Koppa, A., Villanueva, O. B., Tronquo, E., Zhong, F., Beck, H., Hulsman, P., Dorigo, W., Verhoest, N. E., et al. (2024). Gleam4: global land evaporation dataset at 0.1 resolution from 1980 to near present. Journal Placeholder.

Miralles, D. G., Holmes, T. R., De Jeu, R. A., Gash, J. H., Meesters, A. G., and Dolman, A. J. (2011). Global land-surface evaporation estimated from satellite-based observations. Hydrology and Earth System Sciences, 15(2):453–469.

Moutier, W., Bourgeois, Q., Tetzlaff, A., Clerbaux, N., Stöckli, R., Schröder, M., and Hollmann, R. (2024). CM SAF surface radiation and fluxes from meteosat first and second generation edition 1 (landflux ed. 1). DOI:10.5676/EUM_SAF_CM/SLF_METEOSAT/V001

Moutier, W., Clerbaux, N., Gellens-Meulenberghs, F., Arboleda, A., and Barrios, M. (2023a). Validation report for [latent and sensible heat flux climate data record]. Validation Report 1.1, Climate Monitoring Satellite Applications Facility (CM SAF).

Moutier, W., Gellens-Meulenberghs, F., Arboleda, A., Barrios, M., Ghilain, N., and Clerbaux, N. (2023b). Algorithm theoretical basis document for [meteosat latent and sensible heat fluxes]. Technical Report 1.2, Climate Monitoring Satellite Applications Facility (CM SAF).

Muñoz Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., Boussetta, S., Choulga, M., Harrigan, S., Hersbach, H., Martens, B., Miralles, D. G., Piles, M., Rodríguez-Fernández, N. J., Zsoter, E., Buontempo, C., and Thépaut, J.-N. (2021). Era5-land: a state-of-the-art global reanalysis dataset for land applications. Earth System Science Data, 13(9):4349–4383.

Mueller, B. and Seneviratne, S. I. (2014). Systematic land climate and evapotranspiration biases in cmip5 simulations. Geophysical research letters, 41(1):128–134.

Niedorf, A., Konrad, H., Finkensieper, S., Oliveira, R. J., Pondrom, M., Radovan, A., Cloché, S., Fennig, K., Sikorski, T., Roca, R., Schröder, M., and Kidd, C. (2024). Algorithm Theoretical Basis Document - Global Interpolated RAinFall Estimation (GIRAFE v1). Algorithm Theoretical Basis Document SAF/CM/DWD/ATBD/GIRAFE/1.3, EUMETSAT Satellite Application Facility on Climate Monitoring (CM SAF), Deutscher Wetterdienst (DWD), Germany. Version 1.3, Issue Date: 26 January 2024.

Nijzink, R., Almeida, S., Pechlivanidis, I., Capell, R., Gustafssons, D., Arheimer, B., Parajka, J., Freer, J., Han, D., Wagener, T., et al. (2018). Constraining conceptual hydrological models with multiple information sources. Water Resources Research, 54(10):8332–8362.

Oki, T. and Kanae, S. (2006). Global hydrological cycles and world water resources. science, 313(5790):1068–1072.

Oliveira, P. T. S., Nearing, M. A., Moran, M. S., Goodrich, D. C., Wendland, E., and Gupta, H. V.(2014). Trends in water balance components across the brazilian cerrado. Water Resources Research, 50(9):7100–7114.

Palmer, A. R., Fuentes, S., Taylor, D., Macinnis-Ng, C., Zeppel, M., Yunusa, I., and Eamus, D. (2010). Towards a spatial understanding of water use of several land-cover classes: an examination of relationships amongst pre-dawn leaf water potential, vegetation water use, aridity and modis lai. Ecohydrology: Ecosystems, Land and Water Process Interactions, Ecohydrogeomorphology, 3(1):1–10.

Pan, M., Sahoo, A. K., Troy, T. J., Vinukollu, R. K., Sheffield, J., and Wood, E. F. (2012). Multisource estimation of long-term terrestrial water budget for major global river basins. Journal of Climate, 25(9):3191–3206.

Paredes, P., Trigo, I., de Bruin, H., Simões, N., and Pereira, L. S. (2021). Daily grass reference evapotranspiration with meteosat second generation shortwave radiation and reference et products. Agricultural Water Management, 248:106543.

Pascolini-Campbell, M. A., Reager, J. T., and Fisher, J. B. (2020). Grace-based mass conservation as a validation target for basin-scale evapotranspiration in the contiguous united states. Water Resources Research, 56(2):e2019WR026594.

Perez, M., Lombardi, D., Bardino, G., and Vitale, M. (2024). Drought assessment through actual evapotranspiration in mediterranean vegetation dynamics. Ecological Indicators, 166:112359.

Priestley, C. H. B. and Taylor, R. J. (1972). On the assessment of surface heat flux and evaporation using large-scale parameters. Monthly weather review, 100(2):81–92.

Rajib, A., Evenson, G. R., Golden, H. E., and Lane, C. R. (2018). Hydrologic model predictability improves with spatially explicit calibration using remotely sensed evapotranspiration and biophysical parameters. Journal of hydrology, 567:668–683.

Randall, D. A., Wood, R. A., Bony, S., Colman, R., Fichefet, T., Fyfe, J., Kattsov, V., Pitman, A., Shukla, J., Srinivasan, J., et al. (2007). Climate models and their evaluation. In Climate change 2007: The physical science basis. Contribution of Working Group I to the Fourth Assessment Report of the IPCC (FAR), pages 589–662. Cambridge University Press.

Roca, R., Taburet, N., Lorant, E., Chambon, P., Alcoba, M., Brogniez, H., Cloché, S., Dufour, C., Gosset, M., and Guilloteau, C. (2018). Quantifying the contribution of the megha-tropiques mission to the estimation of daily accumulated rainfall in the tropics. Quarterly Journal of the Royal Meteorological Society, 144:49–63.

Rodell, B., Houser, P., Jambor, U., Gottschalck, J., Mitchell, K., Meng, C., Arsenault, K., Cosgrove, B., Radakovich, J., Bosilovich, M., et al. (2004). The global land data assimilation system this powerful new land surface modeling system integrates data from advanced observing systems to support improved forecast model initialization and hydrometeorological investigations, b. am. meteorol. soc., 85, 381–394.

Ruhoff, A., de Andrade, B. C., Laipelt, L., Fleischmann, A. S., Siqueira, V. A., Moreira, A. A., Barbedo, R., Cyganski, G. L., Fernandez, G. M. R., Brêda, J. P. L. F., et al. (2022). Global evapotranspiration datasets assessment using water balance in south america. Remote Sensing, 14(11):2526.

Sakumura, C., Bettadpur, S., and Bruinsma, S. (2014). Ensemble prediction and intercomparison analysis of grace time-variable gravity field models. Geophysical Research Letters, 41(5):1389–1397.

Schulzweida, U. (2023). *CDO: Climate Data Operators User Guide (v2.1.0).*

Sellers, P., Dickinson, R., Randall, D., Betts, A., Hall, F., Berry, J., Collatz, G., Denning, A., Mooney, H., Nobre, C., et al. (1997). Modeling the exchanges of energy, water, and carbon between continents and the atmosphere. Science, 275(5299):502–509.

Sheffield, J., Ferguson, C. R., Troy, T. J., Wood, E. F., and McCabe, M. F. (2009). Closing the terrestrial water budget from satellite remote sensing. Geophysical Research Letters, 36(7).

Sheffield, J., Wood, E. F., Pan, M., Beck, H., Coccia, G., Serrat-Capdevila, A., and Verbist, K. (2018). Satellite remote sensing for water resources management: Potential for supporting sustainable development in data-poor regions. Water Resources Research, 54(12):9724–9758.

Sirisena, T. J. G., Maskey, S., and Ranasinghe, R. (2020). Hydrological model calibration with streamflow and remote sensing based evapotranspiration data in a data poor basin. Remote sensing, 12(22):3768.

Stahl, K., Weiler, M., van Tiel, M., Kohn, I., Hänsler, A., Freudiger, D., Seibert, J., Gerlinger, K., and Moretti, G. (2022). Impact of climate change on the rain, snow and glacier melt components of streamflow of the river rhine and its tributaries. CHR report no. I, 28.

Swann, A. L. and Koven, C. D. (2017). A direct estimate of the seasonal cycle of evapotranspiration over the amazon basin. Journal of Hydrometeorology, 18(8):2173–2185.

Tan, X., Liu, B., Tan, X., and Chen, X. (2022). Long-term water imbalances of watersheds resulting from biases in hydroclimatic data sets for water budget analyses. Water Resources Research, 58(3):e2021WR031209.

Tangdamrongsub, N., Hwang, C., Borak, J. S., Prabnakorn, S., and Han, J. (2021). Optimizing grace/grace-fo data and a priori hydrological knowledge for improved global terrestial water storage component estimates. Journal of Hydrology, 598:126463.

Trenberth, K. E., Fasullo, J. T., and Kiehl, J. (2009). Earth's global energy budget. Bulletin of the American Meteorological Society, 90(3):311–324.

Trigo, I. F., Dacamara, C. C., Viterbo, P., Roujean, J.-L., Olesen, F., Barroso, C., Camacho-de Coca, F., Carrer, D., Freitas, S. C., García-Haro, J., et al. (2011). The satellite application facility for land surface analysis. International Journal of Remote Sensing, 32(10):2725–2744.

Ukkola, A. and Prentice, I. (2013). A worldwide analysis of trends in water-balance evapotranspiration. Hydrology and Earth System Sciences, 17(10):4177–4187.

Unnisa, Z., Govind, A., Lasserre, B., and Marchetti, M. (2023). Water balance trends along climatic variations in the mediterranean basin over the past decades. Water, 15(10):1889.

Formatted: Font: (Default) Arial, 11 pt

Formatted: Justified

Formatted: Font: Times New Roman, 12 pt

Formatted: Left, Space After: 0 pt, Line spacing: single

van den Hurk, B., Viterbo, P., Beljaars, A., and Betts, A. (2000). Offline validation of the era40 surface scheme. Technical report, Citeseer.

Wang, S., Huang, J., Li, J., Rivera, A., McKenney, D. W., and Sheffield, J. (2014a). Assessment of water budget for sixteen large drainage basins in canada. Journal of Hydrology, 512:1–15.

Wang, S., Huang, J., Yang, D., Pavlic, G., and Li, J. (2015). Long-term water budget imbalances and error sources for cold region drainage basins. Hydrological Processes, 29(9):2125–2136.

Wang, S., McKenney, D. W., Shang, J., and Li, J. (2014b). A national-scale assessment of long-term water budget closures for canada's watersheds. Journal of Geophysical Research: Atmospheres, 119(14):8712–8725.

Wang, Z., Zhan, C., Ning, L., and Guo, H. (2021). Evaluation of global terrestrial evapotranspiration in cmip6 models. Theoretical and Applied Climatology, 143:521–531.

Wang-Erlandsson, L., Tobian, A., Van der Ent, R. J., Fetzer, I., te Wierik, S., Porkka, M., Staal, A., Jaramillo, F., Dahlmann, H., Singh, C., et al. (2022). A planetary boundary for green water. Nature Reviews Earth & Environment, 3(6):380–392.

Wong, J. S., Zhang, X., Gharari, S., Shrestha, R. R., Wheater, H. S., and Famiglietti, J. S. (2021). Assessing water balance closure using multiple data assimilation–and remote sensingbased datasets for canada. Journal of Hydrometeorology, 22(6):1569–1589.

Wu, C., Yeh, P. J.-F., Wu, H., Hu, B. X., and Huang, G. (2019). Global analysis of the role of terrestrial water storage in the evapotranspiration estimated from the budyko framework at annual to monthly time scales. Journal of Hydrometeorology, 20(10):2003–2021.

Xiong, J., Xu, L., Chandanpurkar, H. A., Famiglietti, J. S., Zhang, C., Ghiggi, G., Guo, S., Pan, Y., Vishwakarma, B. D., et al. (2023). Et-wb: water-balance-based estimations of terrestrial evaporation over global land and major global basins. Earth System Science Data, 15(10):4571–4597.

Xue, B.-L., Wang, L., Li, X., Yang, K., Chen, D., and Sun, L. (2013). Evaluation of evapotranspiration estimates for two river basins on the tibetan plateau by a water balance method. Journal of Hydrology, 492:290–297.

Yang, Y., Roderick, M. L., Guo, H., Miralles, D. G., Zhang, L., Fatichi, S., Luo, X., Zhang, Y., McVicar, T. R., Tu, Z., et al. (2023). Evapotranspiration on a greening earth. Nature Reviews Earth & Environment, 4(9):626–641.

Zarch, M. A. A., Sivakumar, B., and Sharma, A. (2015). Assessment of global aridity change. Journal of Hydrology, 520:300–313.

Zeng, Z., Piao, S., Lin, X., Yin, G., Peng, S., Ciais, P., and Myneni, R. B. (2012). Global evapotranspiration over the past three decades: estimation based on the water balance equation combined with empirical models. Environmental Research Letters, 7(1):014026.

Zhang, L., Marshall, M., Vrieling, A., and Nelson, A. (2023a). The divergence of energyand water-balance evapotranspiration estimates in humid regions. Journal of Hydrology, 624:129971.

Zhang, X., Duan, J., Cherubini, F., and Ma, Z. (2023b). A global daily evapotranspiration deficit index dataset for quantifying drought severity from 1979 to 2022. Scientific Data, 10(1):824.

Zhang, X., Ren, C., Wang, Z., Li, X., and Zhang, Y. (2025). Gap-filling grace and grace-fo data with a climate adjustment scheme using singular spectrum analysis. Journal of Hydrology, page 132782.

Zhang, Y., Leuning, R., Chiew, F. H., Wang, E., Zhang, L., Liu, C., Sun, F., Peel, M. C., Shen, Y., and Jung, M. (2012). Decadal trends in evaporation from global energy and water balances. Journal of Hydrometeorology, 13(1):379–391.

Zhong, F., Jiang, S., van Dijk, A. I., Ren, L., Schellekens, J., and Miralles, D. G. (2022). Revisiting large-scale interception patterns constrained by a synthesis of global experimental data. Hydrology and Earth System Sciences, 26(21):5647–5667.

7 Appendix

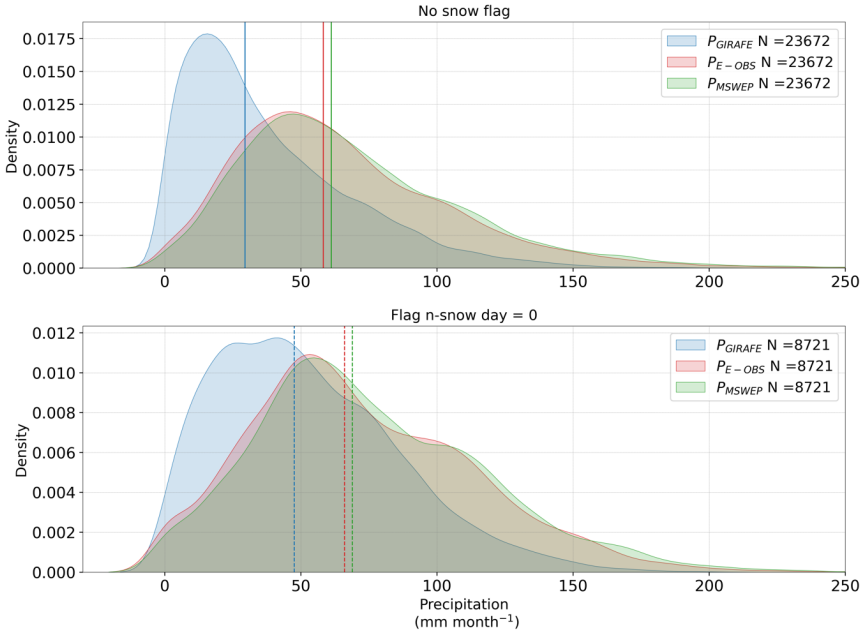


Figure 19: Probability density functions (PDFs) of monthly precipitation (P) estimates for all selected basins over the period 2004–2020, based on three regridded (0.25°) datasets: GIRAFE (blue), MSWEP (green), and E-OBS (red). The three panels correspond to different thresholds on the number of snow days used to filter the GIRAFE data: no snow flag (top) and, and 0 snow days (bottom). The number of matchups (N) varies accordingly. Vertical solid lines indicate the median precipitation values for each dataset.

Deleted: red

Deleted: green

Deleted: , fewer than 20 snow days (middle),

Deleted: fewer than 5

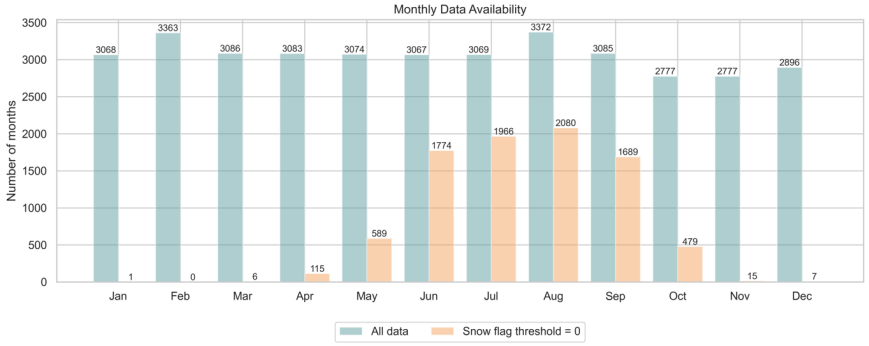


Figure 20: Monthly data availability used in the analysis, showing the number of valid months across all basins for each calendar month. Blue bars represent the full dataset ("All data"), while orange bars correspond to the subset where the GIRAFE snow flag equals zero (i.e., months with no snow-affected grid cells).

Figure 21: Mean annual cycles of hydrological variables over the Düsseldorf basin for the period 2004–2020. The top left panel shows the monthly climatology of river

Formatted: Keep with next

Formatted: Font: 10 pt, Bold

Formatted: Font: 10 pt, Bold

Formatted: Font: 10 pt, Bold

Formatted: Font: (Default) Arial, 10 pt, Bold

Formatted: Justified

Formatted: Font: (Default) Arial, 10 pt, Bold

Formatted: Caption

Formatted: Keep with next

Formatted: Font: Bold

Formatted: Justified

Formatted: Font: Bold

Formatted: Font: Bold

Formatted: Font: (Default) Arial, 11 pt, Bold

discharge (Q), while the top right panel displays the average monthly change in water storage (dS/dt) derived from GRACE data. The middle left panel presents evapotranspiration (ET) estimates from five independent datasets: CM SAF, LSA SAF, GLEAM, ERA5-Land, and GLDAS 2.1. The middle right panel shows monthly precipitation from three datasets: GIRAFE, E-OBS, and MSWEP. The bottom left panel displays ET derived from the water balance approach, calculated using different combinations of precipitation (P) datasets with (ET_{WB}) and without (ET_{PQ}) considering changes in water storage.

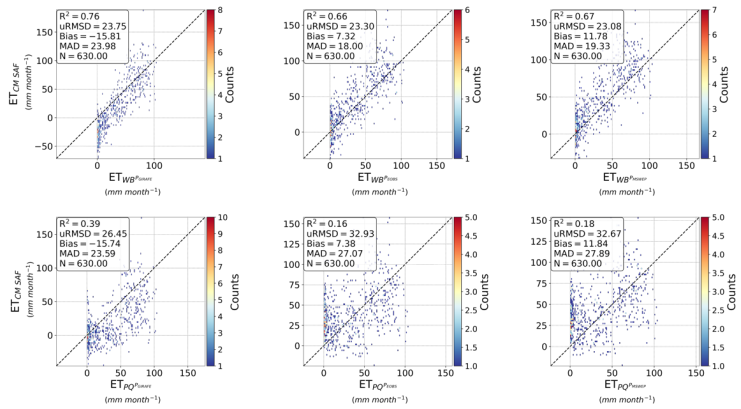


Figure 22: Density scatter plots comparing monthly evapotranspiration estimates from the CM SAF product ($ET_{CM\ SAF}$ in mm month^{-1}) with water balance-derived evapotranspiration estimates with (ET_{WB} in mm month^{-1}; top panels) and without (ET_{PQ} in mm month^{-1}; bottom panels) considering $\frac{dS}{dt}$ for matchups from the six largest river basins from our database. GIRAFE (left panels), E-OBS (middle panels), and MSWEP (right panels) precipitation products are considered for the calculation of ET_{WB} and ET_{PQ}. The color scale indicates point density, with warmer colors representing higher densities. The dashed black line represents the 1:1 relationship. Statistical metrics, including the coefficient of determination (R^2), unbiased root mean square deviation (uRMSD), bias, mean absolute deviation (MAD), and the number of observations (N), are provided in each panel.

Formatted: Font: (Default) Arial, 11 pt, Bold
Formatted: Font: (Default) Arial, 11 pt, Bold
Formatted: Font: Bold
Formatted: Font: Bold
Formatted: Font: Bold
Formatted: Font: (Default) Arial, 11 pt, Bold
Formatted: Caption
Deleted: 20
Deleted: seven largest

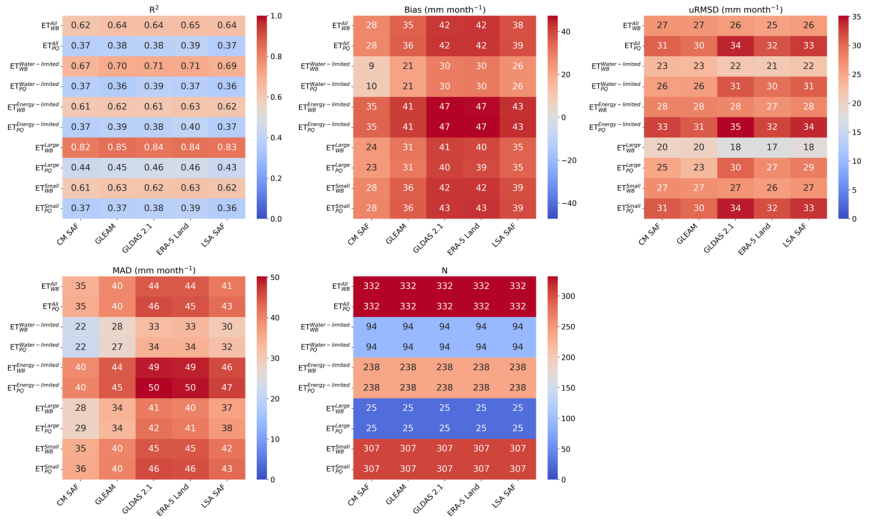


Figure 23: Heatmaps of statistical metrics assessing the agreement between monthly evapotranspiration estimates from different datasets (CM SAF, GLEAM, GLDAS 2.1, ERA5-Land, and LSA SAF) and water balance-derived evapotranspiration estimates considering (ET_{WB}) or not (ET_{PQ}) $\frac{dS}{dt}$ and based on the GIRAFE precipitation product. Statistics are made for each basin and averaged for different subsets: All, water-limited (AI > 1), energy-limited (AI ≤ 1), small (< 90 000 km²) and large (≥ 90 000 km²). The panels display (top-left) the coefficient of determination (R), (top-middle) bias (mm month $^{-1}$), (top-right) unbiased root mean square deviation (uRMSD, mm month^{-1}), (bottom-left) median absolute deviation (MAD, mm month^{-1}), and (bottom-right) the number of observations (N). Warmer colors indicate higher values, while cooler colors represent lower values.

Deleted: 21

Table 2: Summary information of the six largest GRDC river stations (area (≥ 90 000 km²) extracted from the dataset. The columns include the GRDC station number (grdc_no), river and station names, country code (ISO 3166), geographical coordinates (latitude and longitude in decimal degrees), catchment area (in km²), and gauge altitude (in meters above sea level).

gdrc_no	River	Station	Country	Lat (°N)	Lon (°E)	Area (km²)	Altitude (m)
6279500	Prypyats'	Mozyr	BY	52.05	29.27	101000	113.0
6340110	Elbe river	Neu Darchau	DE	53.23	10.89	131950	5.68
6357010	Oder river	Hohensaaten-Finow	DE	52.87	14.14	109564	0.16
64350060	Rhine river	Lobith	NL	51.84	6.11	160800	8.53
6544100	Tisa	Senta	RS	45.93	20.08	140130	74
6742900	Danube river	Ceatal Izmail	RO	45.22	28.72	807000	0.6

Deleted: seven largest

Formatted Table

Deleted: 6955430 ... [6]

8 Glossary – List of Acronyms in alphabetical order

AD	Applicable Document
AGRMET	AGRicultural METeorological modeling system
AI	Aridity Index
CDOP	Continuous Development and Operations Phase
CDR	Climate Data Record
CFC	Cloud Fractional Cover
CM SAF	EUMETSAT Satellite Application Facility on Climate Monitoring
CMIP	Coupled Model Intercomparison Project
CSR	Center for Space Research
DR	Data Record
DWD	Deutscher Wetterdienst
E-OBS	Ensemble of Gridded Observations
ECA&D	European Climate Assessment & Dataset
ECMWF	European Centre for Medium-Range Weather Forecasts
ERA5-Land	ECMWF Reanalysis 5th Generation (Land)
ET	Evapotranspiration
EUMETSAT	European Organisation for the Exploitation of Meteorological Satellites
GEWEX	Global Energy and Water Exchanges
GFZ	GeoForschungsZentrum
GIRAFE	Global Interpolated RainFall Estimation
GLDAS	Global Land Data Assimilation System

GLEAM	Global Land Evaporation Amsterdam Model
GPCP	Global Precipitation Climatology Project
GRACE	Gravity Recovery And Climate Experiment
GRACE-FO	GRACE Follow-On
GRDC	Global Runoff Data Centre
H	Sensible Heat
H-TESSEL	Hydrology Tiled ECMWF Scheme for Surface Exchanges over Land
IPCC	Intergovernmental Panel on Climate Change
IR	Infrared
ISO	International Organization for Standardization
JPL	Jet Propulsion Laboratory
LAI	Leaf Area Index
LE	Latent Heat
LSA SAF	Satellite Application Facility on Land Surface Analysis
LST	Land Surface Temperature
MAD	Mean Absolute Deviation
MMAC	Monthly Mean Annual Cycle
MSWEP	Multi-Source Weighted-Ensemble Precipitation
MVIRI	Meteosat Visible and InfraRed Imager
NASA	National Aeronautics and Space Administration
NRT	Near Real Time
NWP	Numerical Weather Prediction
P	Precipitation
PALS	Protocol for the Analysis of Land Surface models

PDFs	Probability Density Functions
Perc	Percentile
PLUMBER	Land sUrface Model Benchmarking Evaluation pRoject
PMW	Passive Microwave
PT	Priestley-Taylor
Q	Discharge
R^2	Coefficient of Determination
RD	Reference Document
SAF	Satellite Application Facility
SEVIRI	Spinning Enhanced Visible and InfraRed Imager
SM	Soil Moisture
SRB	Surface Radiation Balance
TAPEER	Tropical Amount of Precipitation with an Estimate of ERrors
uRMSD	Unbiased Root Mean Square Deviation

Page 9: [1] Deleted	**William Moutier**	**10/9/25 4:17:00 PM**
Page 31: [2] Deleted	**William Moutier**	**9/16/25 3:38:00 PM**
Page 31: [3] Deleted	**William Moutier**	**9/17/25 10:22:00 AM**
Page 32: [4] Deleted	**William Moutier**	**9/17/25 10:31:00 AM**
Page 32: [5] Deleted	**William Moutier**	**9/17/25 10:27:00 AM**
Page 58: [6] Deleted	**William Moutier**	**9/15/25 11:39:00 AM**