Journal of Hydrology: Regional Studies 63 (2026) 103062

Contents lists available at ScienceDirect

Journal of Hydrology: Regional Studies

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

# Evaluating global precipitation datasets over Sicily: From daily estimates to extreme events

Mehmet Berkant Yıldız, Fabio Di Nunno *, Giovanni de Marinis, Francesco Granata

*University of Cassino and Southern Lazio, Department of Civil and Mechanical Engineering (DICEM), Via Di Biasio, 43, Frosinone, Cassino 03043, Italy*

ARTICLE INFO

*Keywords:*
Precipitation
Extreme events
GEV
POT
Sicily

ABSTRACT

*Study region:* Sicily, the largest island in the Mediterranean, exhibits complex hydro-climatic variability due to its diverse topography and proximity to contrasting maritime and continental influences. Accurate precipitation data are therefore essential for reliable hydrological modeling, disaster risk management, and climate-related assessments in this region.

*Study focus:* This study evaluates the performance of 11 widely used global daily precipitation datasets, satellite-based (GPM IMERG, TRMM, PERSIANN family), reanalysis (ERA5-Land, GLDAS-2), and blended products (MSWEP, HydroGFD), against ground-based observations across Sicily for the 2003–2023 period. A suite of statistical metrics (MAE, RMSE, KGE, Pearson's *r*) was used to assess accuracy, precision, and bias. Additionally, the capability of MSWEP to represent extremes was analyzed using the Generalized Extreme Value (GEV) distribution and Peaks Over Threshold (POT) method.

*New hydrological insights for the region:* MSWEP consistently delivered the highest overall accuracy, followed by ERA5-Land and HydroGFD. All datasets effectively captured general precipitation patterns but struggled to reproduce high-intensity events, particularly in mountainous areas such as the northeastern slopes of Mount Etna. While MSWEP accurately represented moderate extremes, it underestimated the magnitude and frequency of severe events. These results emphasize the importance of local calibration and validation when applying global datasets in hydroclimatically complex regions. Although the 2003–2023 record limits long-term climatic inference, the findings provide practical guidance for short-term hydrological studies and dataset selection in Mediterranean settings, contributing to improved flood assessment and water resource planning.

## 1. Introduction

In an era marked by increasingly frequent and intense hydroclimatic extremes, the availability of accurate daily precipitation data has become indispensable for advancing research and decision-making in hydrology, climate science, agricultural planning, and disaster risk management. Reliable and spatially coherent precipitation measurements are crucial for ensuring sustainable water resource management, anticipating and mitigating flood and drought hazards, and improving the fidelity of climate impact

* Corresponding author.
*E-mail addresses:* mehmetberkant.yildiz@unicas.it (M.B. Yıldız), fabio.dinunno@unicas.it (F. Di Nunno), demarinis@unicas.it (G. de Marinis), f.granata@unicas.it (F. Granata).

https://doi.org/10.1016/j.ejrh.2025.103062
Received 7 August 2025; Received in revised form 29 November 2025; Accepted 16 December 2025

assessments (Akbas and Ozdemir, 2024). Yet, the stochastic nature of atmospheric processes poses profound challenges to the consistent monitoring of precipitation across time and space, particularly in heterogeneous and data-scarce regions (Flitcroft et al., 1989; Lu et al., 2020). These complexities underscore the growing scientific and operational imperative to develop continuous, high-resolution precipitation datasets that can support both long-term climate analyses and short-term hazard preparedness. Rain gauges have traditionally served as the primary instruments for direct precipitation measurement, offering high temporal resolution and ground-truth accuracy. Nonetheless, their effectiveness may be limited in regions where the monitoring network is sparse, unevenly distributed, or inadequately maintained. In such cases, spatial interpolation becomes necessary to estimate rainfall across unsampled areas; however, this process introduces significant uncertainty, particularly when high spatial resolution is required. This limitation becomes particularly evident in areas with highly variable microclimates, where rainfall can vary significantly over short distances, variability that isolated point measurements may fail to capture (Lei et al., 2021; Rao et al., 2024). Moreover, the precision of gauge-based observations can be influenced by several physical factors, including wind-induced undercatch (Nešpor and Sevruk, 1999; Pollock et al., 2018), evaporation losses from the collecting surface (Leeper and Kochendorfer, 2015), and aerodynamic inefficiencies in instrument design (Cauteruccio et al., 2024). These sources of uncertainty, especially under challenging meteorological conditions, may reduce the representativeness of conventional observations and highlight the importance of integrating alternative or complementary data sources.

Satellite-based and reanalysis datasets have become essential for addressing these gaps, offering global-scale precipitation information (Du et al., 2022; Cheng et al., 2025). Satellite sensors deliver homogeneous, continuous data by frequently monitoring the Earth's surface using infrared, microwave, and radar technologies. Global missions such as TRMM and GPM have notably expanded precipitation monitoring, particularly in tropical and subtropical regions (Hou et al., 2014; Skofronick-Jackson et al., 2017). Similarly, reanalysis datasets, generated through numerical weather prediction models, reconstruct historical atmospheric conditions with temporal consistency.

Despite their broad coverage and high temporal resolution, these datasets often face challenges in accurately representing precipitation in regions characterized by highly variable microclimates. In particular, they may systematically underestimate extreme and localized high-intensity precipitation events. These limitations become especially pronounced in Mediterranean regions, where sharp climatic gradients amplify spatial variability in rainfall. Accurate, high-resolution precipitation datasets are therefore essential to effectively capture hydroclimatic variability in such environments.

Although satellite- and reanalysis-based precipitation datasets have been widely applied across the Mediterranean, focused studies on Sicily remain limited, particularly those offering systematic, high-resolution comparisons across a broad suite of global products (Lo Conti et al., 2012). Existing research has largely emphasized seasonal or annual scales, often overlooking daily precipitation performance, which is critical for operational hydrology and disaster risk management. Moreover, only a few studies (e.g., Maugeri et al., 2015; Noto et al., 2022) have investigated the ability of these datasets to capture extreme rainfall events, despite Sicily's known vulnerability to high-intensity storms.

This study addresses these gaps through a comprehensive evaluation of 11 widely used daily precipitation datasets over Sicily, including satellite-based (e.g., GPM IMERG, TRMM), reanalysis (e.g., ERA5-Land, GLDAS), blended products (e.g., MSWEP, HydroGFD), and observational gridded datasets (E-OBS). To guide the analysis, the study attempts to answer the following two questions:

(1) Which global precipitation datasets most accurately reproduce daily rainfall over Sicily compared to ground-based observations?
(2) How well do these datasets represent extreme precipitation, including events associated with major floods?

The evaluation combines a multi-metric assessment of daily performance (e.g., RMSE, MAE) with a focused extreme-value analysis using the GEV distribution and the POT method. This two-tiered approach allows us to assess both general accuracy and the ability of datasets to reproduce historical extreme rainfall events at the local scale.

Compared to existing literature, the study offers three distinctive contributions:

(i) the first island-wide intercomparison of 11 global daily precipitation datasets over Sicily, encompassing satellite, reanalysis, blended, and observational products;
(ii) a novel linkage between general performance metrics and formal extreme-value analysis, offering insight into how datasets capture extreme precipitation events and complementing earlier Mediterranean studies (Mascaro et al., 2018; Caroletti et al., 2019; Akbas and Ozdemir, 2024);
(iii) spatially explicit assessments of dataset performance across Sicily's diverse microclimatic conditions.

To our knowledge, this is the first study to integrate daily accuracy assessments with extreme-event analysis across the entire island. By doing so, it provides new regional evidence on the strengths and limitations of global precipitation datasets in representing both ordinary and flood-prone rainfall events.

Overall, addressing these questions provides scientific and practical insights into the reliability of global precipitation products for hydrological modeling and extreme-event assessment in Mediterranean regions characterized by pronounced hydroclimatic variability.

## 2. Materials and methods

### 2.1. Study area and dataset

Sicily, the largest island in the Mediterranean Sea, lies near 37.5°N and 14°E and spans approximately 25,700 km² just south of mainland Italy ( Fig. 1). The island exhibits a typical Mediterranean climate, characterized by hot, arid summers and mild, wet winters. Annual precipitation across the island ranges from 400 mm to 1400 mm, with the southern and southeastern coasts experiencing semi-arid conditions (<500 mm), while the northern and inland mountainous regions receive over 1000 mm annually (Caracciolo et al., 2018). During summer, monthly totals often fall below 10–20 mm, leading to severe droughts. This pronounced seasonality negatively impacts agricultural sustainability and heightens water demand (Dallan et al., 2025). Conversely, Sicily has been frequently affected by heavy rainfall and flood events over the past two decades. In November 2018, over 300 mm of rain fell across eastern and southern Sicily, causing severe flooding and fatalities (Davolio et al., 2020; Scicchitano et al., 2021), while the October 2021 medicane brought more than 150 mm to Catania in a few hours, inundating the city (Polaris, 2025).

With a population of approximately five million concentrated largely along the coast, Sicily has become increasingly vulnerable to hydroclimatic extremes. Prolonged droughts in the early 2000s and again in 2023–2024 have stressed water resources and agricultural systems. Conversely, extreme rainfall events, including those in 2009, 2018, and 2021, have triggered flash floods and inflicted severe infrastructural damage, underscoring the need for accurate precipitation monitoring in a region marked by pronounced climatic gradients.

In this study, 11 daily precipitation datasets were employed to assess precipitation variability over Sicily using a comprehensive evaluation framework based on a wide range of statistical performance metrics (Table 1). To improve readability and avoid redundancy, the dataset abbreviations listed in Table 1 are used throughout the manuscript. The selected datasets originate from diverse methodological backgrounds and data assimilation strategies, reflecting a variety of observational systems and modeling techniques

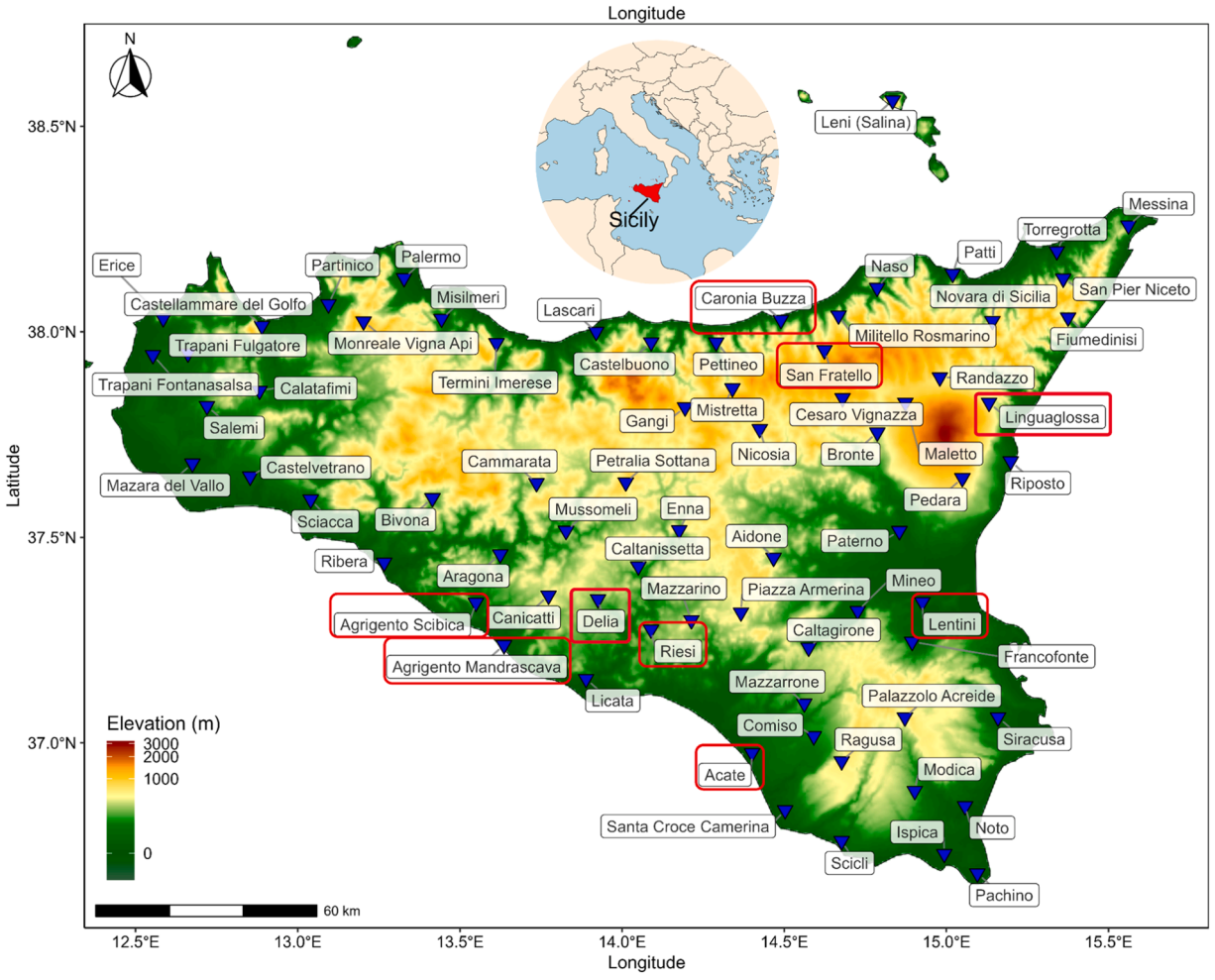


**Fig. 1.** Location of the monitoring stations in Sicily.

**Table 1**
Description of the precipitation dataset.

| Name-short name | Abbreviation | Spatial resolution (degree) | Temporal coverage | Data sources | References |
|---|---|---|---|---|---|
| European gridded observational dataset | E-OBS | 0.10 × 0.10 | January 1950 to present | Gauge | Haylock et al. (2008); ECMWF 2017; Cornes et al. (2018) |
| The Integrated Multi-Satellite Retrievals for GPM Final Precipitation L3 1 day (GPM_3IMERGDF) | GPM_3IMERGDF | 0.10 × 0.10 | June 2000 to present | Satellite + Gauge | Huffman et al. (2019) |
| Tropical Rainfall Measuring Mission / Multi-satellite Precipitation Analysis 3B42 Version 7 | TRMM | 0.25 × 0.25 | March 2000 to February 2020 | Satellite + Gauge | Huffman et al. (2007); Huffman and Bolvin, (2018) |
| Precipitation Estimation from Remotely Sensed Information using Artificial Neural Networks | PERSIANN/PERS | 0.25 × 0.25 | March 2000 to present | Satellite | Hsu et al. (1997); Nguyen et al. (2019) |
| PERSIANN-Cloud Classification System | PERSIANN/CCS | 0.04 × 0.04 | January 2003 to present | Satellite | Hong et al. (2004); Nguyen et al. (2019) |
| PERSIANN-Climate Data Record | PERSIANN/CDR | 0.25 × 0.25 | January 1983 to present | Satellite | Ashouri et al. (2015); Nguyen et al. (2019) |
| PERSIANN-Dynamic Infrared Rain Rate near real-time | PERSIANN/PDIR-Now | 0.04 × 0.04 | March 2000 to present | Satellite | Nguyen et al. (2020a); Nguyen et al. (2020b) |
| ECMWF Reanalysis 5th Generation for Lands | ERA5-Land | 0.10 × 0.10 | January 1950 to present | Reanalysis | Hersbach et al. (2020); Muñoz-Sabater et al. (2021) |
| NASA Global Land Data Assimilation System Version 2 | GLDAS-2 | 0.25 × 0.25 | January 1948 to December 2014 | Reanalysis | Li et al. (2019) |
| Multi-Source Weighted-Ensemble Precipitation, version 2 | MSWEP V2 | 0.10 × 0.10 | January 1979 to present | Reanalysis + Satellite + Gauge | Beck et al. (2017), (2019) |
| Hydrological Global Forcing Data | HydroGFD3.0 | 0.25 × 0.25 | February 1979 to December 2019 | Reanalysis + Satellite | Berg et al. (2018), (2021) |

designed to capture spatial precipitation variability across regions with heterogeneous climatic conditions. Gauge-based gridded datasets are generated by interpolating in-situ station data using statistical or geostatistical methods. In contrast, satellite-based products such as IMERG and TRMM estimate precipitation using remote sensing technologies, mainly passive microwave (PMW) and infrared (IR) sensors, with subsequent bias correction applied through available ground-based observations (Huffman et al., 2019). Reanalysis products like ERA5-Land and GLDAS simulate atmospheric processes by assimilating observational data into physical models based on numerical weather prediction systems (Hersbach et al., 2020). Fused or multi-source products such as MSWEP and HydroGFD aim to enhance accuracy and temporal consistency by combining multiple inputs, including satellite retrievals, ground observations, and reanalysis outputs through sophisticated data merging techniques.

As is well recognized, all precipitation datasets are inherently affected by varying degrees of uncertainty, stemming from differences in sensor capabilities, interpolation methodologies, model architectures, and the density or availability of input data, particularly in regions characterized by sparse observational coverage. Given the multitude of global products currently available, this study focused exclusively on datasets satisfying a set of stringent inclusion criteria: (i) availability of daily temporal resolution, (ii) complete spatial coverage over Sicily, (iii) temporal consistency throughout the common study period 2003–2023, and (iv) a spatial resolution equal to 0.04° × 0.04° (1679 cells covering Sicily), 0.10° × 0.10° (358 cells) and 0.25° × 0.25° (79 cells). Datasets that did not meet these requirements, due to gaps in temporal continuity, coarse resolution, or limited compatibility with observational benchmarks, were excluded from the analysis. The evaluation approach involved matching each observational station to the nearest grid cell in each dataset rather than resampling data to a uniform grid resolution. This method preserves the original spatial detail of each dataset and avoids artificial smoothing effects. Consequently, datasets with spatial resolution finer than 0.25° × 0.25° were preferred, as they better capture localized precipitation variability, especially in Sicily's microclimatic diversity.

There are two general approaches to validating gridded Global Precipitation Datasets (GPDs) against station observations. In the point-to-pixel approach, the precipitation value of each grid cell is directly compared with that measured at the gauge (i.e. point-based observations are matched to the corresponding pixel, Dinku et al., 2008; Keikhosravi-Kiany et al., 2022). In the alternative pixel-to-pixel approach, the station data are first interpolated onto a regular grid (e.g. using kriging or inverse-distance weighting) with a resolution similar to that of the satellite or reanalysis product, and the comparison is then carried out between two gridded fields. In this study, the point-to-pixel approach was adopted. Each station of the Servizio Informativo Agrometeorologico Siciliano (SIAS) [http://www.sias.regione.sicilia.it] was assigned to the grid cell whose center lies closest to the station location in each GPD. This method was preferred because the main objective is to evaluate how accurately the gridded products reproduce local precipitation as observed at the SIAS gauges, rather than to compare them with an interpolated precipitation field. Interpolating station data to create a gridded reference (pixel-to-pixel) would introduce additional smoothing and interpolation errors, especially given the heterogeneous spatial distribution of stations and the complex topography of Sicily. By maintaining the original point-based nature of the reference data, the analysis remains more faithful to actual ground conditions and avoids introducing extra sources of uncertainty.

A key strength of this study lies in the use of these high-resolution, ground-based precipitation measurements, derived from a dense and spatially representative network of SIAS meteorological stations distributed across the island, which served as the empirical reference standard for assessing the accuracy and reliability of each gridded precipitation product.

### 2.2. Evaluation metrics

The error and performance metrics used in this study were selected to quantitatively assess the agreement between measured (station-based) data and gridded datasets. The mathematical expressions, formulae, and value ranges for interpreting these metrics are summarized in Table 2. Continuous assessment metrics rely on traditional statistical measures that evaluate the entire time series of precipitation data. In these methods, observed precipitation (O) is compared with estimated values derived from reanalysis, satellite, merged, or gridded in situ datasets (E). The metrics listed in Table 2 were implemented using the *hydroGOF* package in R (Zambrano-Bigiarini, 2014; Akbas and Ozdemir, 2024). In total, nine statistical indicators were employed in this study: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Percent Bias (PBIAS), Pearson correlation coefficient (r), Nash–Sutcliffe Efficiency (NSE), Modified NSE (mNSE), Index of Agreement (d), Modified Index of Agreement (md), and Kling–Gupta Efficiency (KGE). These metrics were chosen to provide a multidimensional assessment of model performance, accounting for accuracy, precision, variability, and temporal correlation. Although the study period (2003–2023) spans only 20 years, it includes over 7000 daily observations per location, providing a statistically robust basis for evaluating dataset performance at the daily timescale. This specific period was selected because it represents the maximum overlap across all gridded datasets and ensures data consistency in terms of sensor quality and reanalysis techniques. Notably, the aim of the study is not to establish climatological normals, but to evaluate how well each dataset replicates observed daily precipitation dynamics, including extremes. Therefore, the use of these performance metrics allows for a comprehensive and technically sound intercomparison across different precipitation products.

where: $O_i$ = measured value at timestep $i$, $E_i$ = gridded value at timestep $i$, $\overline{O}$ = mean of measured value, $\sigma_o$, $\sigma_E$ = standard deviation of measured and gridded values, respectively, $\alpha = \frac{\sigma_E}{\sigma_o}$(ratio of standard deviations for KGE), $\beta = \frac{\mu_E}{\mu_o}$(ratio of means for KGE), $n$ = total number of observations, $\mathrm{Cov}(O,E)$ = covariance between measured and gridded values.

### 2.3. GEV model and POT method

This study examines the statistical behavior of extreme precipitation events using two widely recognized extreme value analysis techniques: the GEV model (Schellander et al., 2019; Gentilucci et al., 2023) and the POT method (Acero et al., 2011; Agilan et al., 2021). Both methods are applied to observed and modeled precipitation datasets.

The GEV model characterizes the distribution of block maxima (e.g., annual or seasonal maximum precipitation values). It assumes that the maximum values within each block are realizations from a common extreme value distribution. The cumulative distribution function (CDF) of the GEV model is defined as:

$$F(x) = \exp\left\{ -\left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi} \right\} \tag{10}$$

**Table 2**
The continuous metrics used in this study.

| Metric name | Generalized equations | | Unit / range / best value | References |
|---|---|---|---|---|
| Index of agreement ($d$) | $d = 1 - \frac{\sum_{i=1}^{n}(O_i - E_i)^2}{\sum_{i=1}^{n}(\lvert E_i - \overline{O}\rvert + \lvert O_i - \overline{O}\rvert)^2}$ | (1) | [−] / [0,1] / [1] | Willmott, (1981) |
| Kling-Gupta Efficiency (KGE) | $KGE = 1 - \sqrt{(1-r)^2 + (\alpha-1)^2 + (\beta-1)^2}$ | (2) | [−] / [−∞ ,1] / [1] | Gupta et al. (2009) |
| Mean Absolute Error (MAE) | $MAE = \frac{1}{n}\sum_{i=1}^{n}\lvert E_i - O_i\rvert$ | (3) | [mm] / [0,∞ ] / [0] | Wilks, (2011); Zambrano-Bigiarini 2017 |
| Modified index of agreement (md) | $md = 1 - \frac{\sum_{i=1}^{n}\lvert O_i - E_i\rvert}{\sum_{i=1}^{n}(\lvert E_i - \overline{O}\rvert + \lvert O_i - \overline{O}\rvert)}$ | (4) | [−] / [0,1] / [1] | Willmott, (1981) |
| Modified Nash-Sutcliffe Efficiency (mNSE) | $mNSE = 1 - \frac{\sum_{i=1}^{n}\lvert E_i - O_i\rvert}{\sum_{i=1}^{n}\lvert O_i - \overline{O}\rvert}$ | (5) | [−] / [−∞ ,1] / [1] | Krause, Boyle, and Bäse, (2005) |
| Nash-Sutcliffe Efficiency (NSE) | $NSE = 1 - \frac{\sum_{i=1}^{n}(E_i - O_i)^2}{\sum_{i=1}^{n}(O_i - \overline{O})^2}$ | (6) | [−] / [−∞ ,1] / [1] | Nash and Sutcliffe, (1970); Krause et al. (2005) |
| Percent Bias (PBIAS) | $PBIAS = 100 * \frac{\sum_{i=1}^{n}(E_i - O_i)}{\sum_{i=1}^{n} O_i}$ | (7) | [%] / [0,∞ ] / [0] | Sorooshian et al. (1993); Moriasi et al. (2007) |
| Pearson correlation coefficient ($r$) | $r = \frac{Cov(O,E)}{\sigma_o \sigma_E}$ | (8) | [−] / [−1,1] / [1] | Zambrano-Bigiarini 2014 |
| Root Mean Square Error (RMSE) | $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(E_i - O_i)^2}$ | (9) | [mm] / [0,∞ ] / [0] | Wilks, (2011); Zambrano-Bigiarini 2014 |

with the domain constraint $1+\xi\frac{x-\mu}{\sigma}>0$, where:

- μ is the location parameter,
- σ> 0 is the scale parameter,
- ξ is the shape parameter.

After extracting block maxima from the dataset (e.g., annual maximum precipitation), the GEV parameters are estimated using the maximum likelihood estimation (MLE) method. Model performance is assessed using criteria such as the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). Additionally, return levels (or quantiles) for specified return periods (e.g., 10, 25, 50, 75, and 100 years) are calculated based on the fitted GEV model to evaluate the severity of extreme events. The return level $Z_T$ corresponding to return period T is obtained as:

$$Z_T\begin{cases}\mu-\frac{\sigma}{\xi}\left[1-\left\{-\log\left(1-\frac{1}{T}\right)\right\}^{-\xi}\right], \xi\neq 0\\ \mu-\sigma\log\left(1-\frac{1}{T}\right), \xi=0\end{cases} \tag{11}$$

In the POT method, all precipitation values exceeding a high threshold u (e.g., the 90th, 95th, or 99th percentiles) are analyzed. Let Y=X − u denote the excess over the threshold. Under suitable regularity conditions, the distribution of Y can be modeled by the Generalized Pareto Distribution (GPD) with the following cumulative distribution function (CDF):

$$F(y)=1-\left(1+\frac{\xi y}{\sigma}\right)^{-\frac{1}{\xi}},\quad y>0,\quad 1+\frac{\xi y}{\sigma} \tag{12}$$

where:

- σ> 0 is the scale parameter,
- ξ is the shape parameter.

The parameters σ and $\xi$ are estimated using the maximum likelihood estimation (MLE) method based solely on the exceedance values. Model uncertainty is assessed by calculating 95 % confidence intervals for the estimated parameters (e.g., using the profile likelihood method via the ci.fevd() function with type = "parameter" in the *extRemes* package).

The analysis was carried out separately for observed and modeled precipitation datasets. For the GEV model, annual maximum precipitation values were extracted to represent block maxima, and the distribution was fitted using maximum likelihood estimation. Model performance was assessed using information criteria such as AIC and BIC, and return levels were calculated for selected return

**Table 3**
Ranking of the precipitation dataset. Colorbar ranges from green (best rank) to white (worst rank).

| Ranks | MAE | RMSE | r | d | KGE | md | mNSE | NSE | PBIAS | | Final Ranks |
|---|---|---|---|---|---|---|---|---|---|---|---|
| E-OBS | 4 | 4 | 5 | 8 | 7 | 5 | 4 | 4 | 9 | MSWEP | 1 |
| ERA5-Land | 2 | 2 | 2 | 2 | 1 | 2 | 2 | 2 | 3 | ERA5-Land | 2 |
| GLDAS | 5 | 6 | 8 | 10 | 9 | 6 | 5 | 5 | 10 | HydroGFD | 3 |
| GPM-IMERG | 6 | 5 | 4 | 4 | 4 | 4 | 6 | 6 | 1 | GPM-IMERG | 4 |
| HydroGFD | 3 | 3 | 3 | 3 | 2 | 3 | 3 | 3 | 2 | E-OBS | 5 |
| MSWEP | 1 | 1 | 1 | 1 | 3 | 1 | 1 | 1 | 6 | GLDAS | 6 |
| PERSIANN | 11 | 11 | 10 | 11 | 11 | 11 | 10 | 11 | 11 | PERSIANN/CDR | 7 |
| PERSIANN/CDR | 8 | 9 | 7 | 6 | 5 | 8 | 8 | 9 | 5 | PERSIANN | 8 |
| PERSIANN/CSS | 7 | 7 | 9 | 7 | 6 | 7 | 7 | 7 | 7 | PERSIANN/PDIR | 9 |
| PERSIANN/PDIR | 9 | 10 | 6 | 5 | 8 | 9 | 9 | 10 | 8 | TRMM | 10 |
| TRMM | 10 | 8 | 11 | 9 | 10 | 10 | 11 | 8 | 4 | PERSIANN/CSS | 11 |

periods.

For the POT method, thresholds corresponding to the 90th, 95th, and 99th percentiles were chosen to identify extreme precipitation events. For each threshold, exceedances (i.e., values above the threshold) were extracted and modeled using the GPD. Parameter estimates, along with their 95 % confidence intervals, were used to evaluate model reliability and uncertainty. Results from both methods were compared by focusing on the best-performing model and examining the stations with the highest and lowest performance, using both observed and modeled data to evaluate how effectively the approaches captured extreme precipitation characteristics.

## 3. Results

### 3.1. Ranking precipitation datasets based on comparison with observed data from monitoring stations

The datasets were evaluated based on their ability to represent daily precipitation values in Sicily. Performance was assessed using a range of error metrics, including MAE, RMSE, r, d, KGE, md, mNSE, NSE and PBIAS. For each dataset, these metrics were calculated across all stations, and average values were obtained. Based on these averages, rankings were established for each metric (Table 3), and an overall ranking was derived by averaging the ranks across all error metrics. All evaluations used in situ station observations as the reference standard.

In the overall assessment, the MSWEP, ERA5-Land, and HydroGFD datasets demonstrated clearly superior performance compared to the others. These three products share key strengths, notably their effective integration of diverse data sources and enhanced capacity to represent complex hydroclimatic processes. In contrast, satellite-based datasets generally performed less well. Specifically, satellite-derived products such as PERSIANN, TRMM, and PERSIANN/CSS showed consistently poor performance across nearly all metrics, proving inadequate in representing daily precipitation values. This weaker performance likely stems from their limitations in accurately capturing atmospheric conditions and in modeling local-scale meteorological events with sufficient precision.

Among the datasets analyzed, MSWEP ranked first overall, delivering the highest performance across key error metrics. It achieved top results in critical measures such as MAE, r, d, and RMSE, providing the closest match to in-situ station observations for daily precipitation estimates. However, MSWEP performed relatively poorly in PBIAS, ranking sixth. This weaker performance is primarily attributed to systematic biases arising from uncertainties inherent in the satellite and reanalysis data integrated into the MSWEP framework. Variations in data calibration procedures and spatio-temporal averaging may also contribute to these biases, affecting the accuracy of cumulative precipitation estimates. Despite MSWEP's strong overall performance, these systematic biases highlight areas for potential improvement in data processing, calibration methods, or model structure.

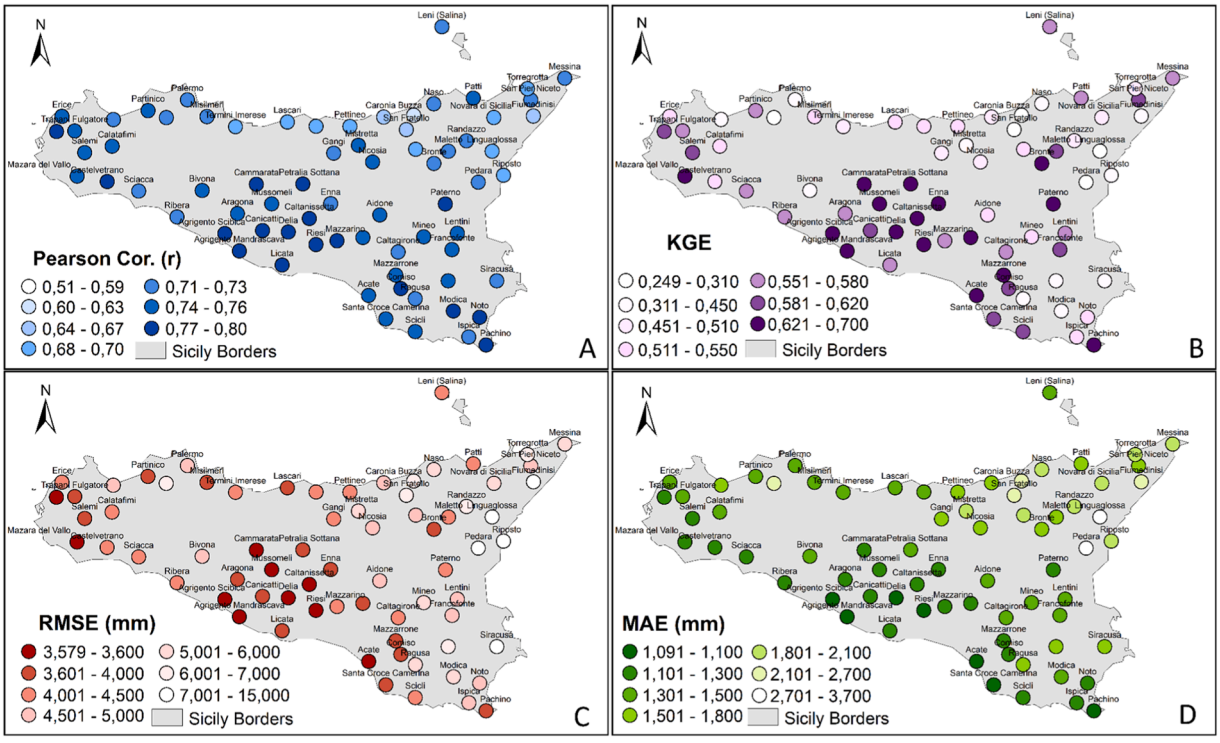


**Fig. 2.** Map of Sicily with r (A), KGE (B), RMSE (C) and MAE (D) computed by comparing observed precipitation at monitoring stations with MSWEP.

The ERA5-Land reanalysis product ranked second, delivering generally consistent and highly accurate results. It demonstrated strength in capturing the variability and timing of precipitation events, achieving the best performance in terms of KGE and RMSE. The superior KGE performance of ERA5-Land likely reflects its enhanced ability to represent extreme precipitation distributions, effectively reducing bias and improving correlation with observed values. Its strong overall performance can be attributed to the detailed representation of atmospheric processes, a critical factors influencing precipitation patterns, especially over regions with highly variable microclimates.

Fig. 2 illustrates the spatial distribution of four key performance metrics, *r* (A), KGE (B), RMSE (C), and MAE (D), for MSWEP daily precipitation estimates compared with in-situ observations. Additional metrics are reported in Figure S1 (supplementary material). Correlation values (*r*, Fig. 2A) range from 0.64 to 0.80, with the highest values (0.74–0.80) in southern and central-southern Sicily, particularly around Delia, Caltanissetta, Trapani, and Paternò, and lower values (0.64–0.70) in the central-northern and northeastern coastal regions, including Militello, Fiumedinisi, Caronia Buzza, and Lascari. KGE (Fig. 2B) shows elevated efficiency in the southern and southeastern regions, reflecting good agreement in both precipitation magnitude and timing. RMSE (Fig. 2C) and MAE (Fig. 2D) indicate the magnitude of errors, with smaller errors in southern Sicily and higher errors in northwestern and northeastern areas, notably around Mistretta, Linguaglossa, and Salemi. These patterns highlight MSWEP's stronger reliability in southern and central-southern areas and reveal reduced accuracy in northern and coastal zones.

Fig. 3 presents the spatial distribution for ERA5-Land (additional metrics in Figure S2). Correlation values (*r*, Fig. 3A) range from 0.60 to 0.78, with higher values in southern stations (Caltanissetta, Ragusa, Agrigento) and lower values in northern coastal regions (Militello, Fiumedinisi, Lascari). KGE (Fig. 3B) indicates higher efficiency in southeastern and central-southern areas, although slightly lower than MSWEP in certain locations. RMSE (Fig. 3C) and MAE (Fig. 3D) are lowest in southern stations and increase in northwestern and far-northern stations (Salemi, Mistretta). Overall, ERA5-Land shows complementary spatial patterns to MSWEP, with consistent efficiency in southern and central-southern Sicily, and slightly lower correlations elsewhere.

Fig. 4 illustrates the spatial distribution for HydroGFD (additional metrics in Figure S3). Strongest correlations ($r \approx 0.64$–$0.70$, Fig. 4A) and highest KGE values (Fig. 4B) are observed in southeastern, eastern, and central Sicily, particularly near Delia, Agrigento Scibica, and Scicli. RMSE and MAE (Figs. 4C and 4D) are lowest in these areas. In contrast, northwestern and far-northern stations, including Salemi, Mistretta, and Linguaglossa, exhibit lower r values and higher errors, highlighting challenges in representing precipitation under complex coastal conditions. The overall spatial patterns confirm that HydroGFD maintains stable performance in many stations while identifying regions with higher uncertainty.

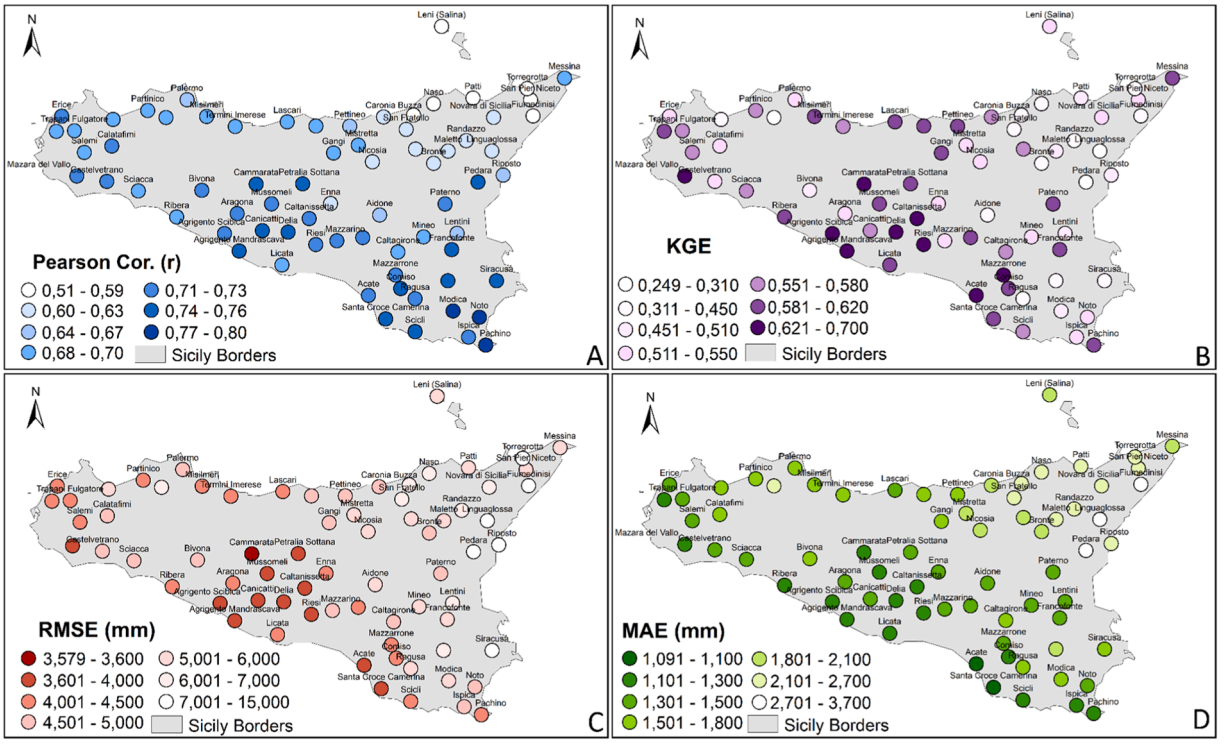


**Fig. 3.** Map of Sicily with r (A), KGE (B), RMSE (C) and MAE (D) computed by comparing observed precipitation at monitoring stations with ERA5-Land.

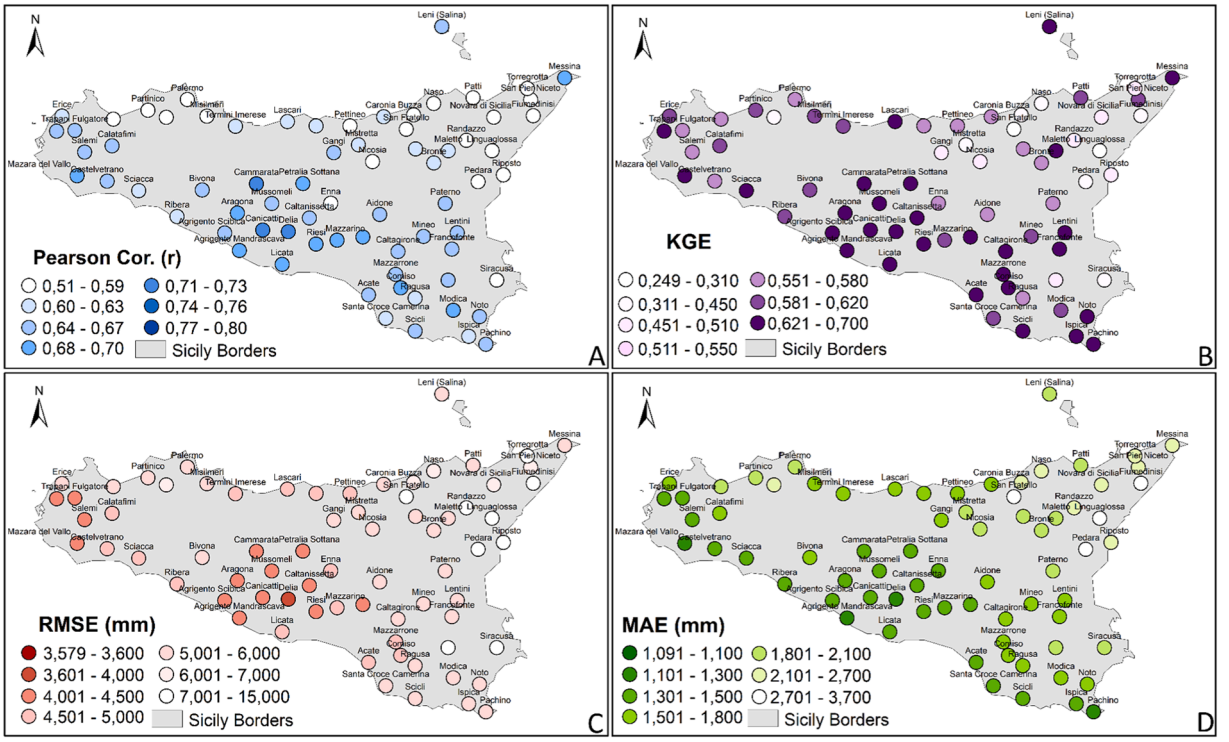


**Fig. 4.** Map of Sicily with r (A), KGE (B), RMSE (C) and MAE (D) computed by comparing observed precipitation at monitoring stations with HydroGFD.

### *3.2. Spatial variability and extreme precipitation analysis: insights from GEV and POT methods*

To further explore the reasons behind the geographical pattern in model accuracy, two monitoring stations were selected as reference points for comparison between observed data and the MSWEP precipitation dataset: Delia and Linguaglossa (highlighted in red in Fig. 1), representing the highest and lowest accuracy levels, respectively. Delia is located in the southern part of Sicily, in the province of Caltanissetta, an area that showed excellent agreement between the precipitation dataset and the monitoring station data. In contrast, Linguaglossa is situated in the northeastern part of Sicily, in the province of Catania, one of the areas with the weakest agreement between the precipitation dataset and station measurements.

The time series and scatterplot for the Delia station (Fig. 5) show a strong agreement between MSWEP estimates and observed precipitation over the analysis period. The correlation coefficient ($r = 0.79$) indicates a robust linear relationship, supported by the tight clustering of points along the 1:1 reference line in the scatterplot. Most precipitation events are well captured in both magnitude and timing, particularly moderate and low-intensity events. However, during high-precipitation episodes (e.g., peaks above 60 mm), MSWEP tends to slightly underestimate the observed values. This pattern is evident in the time series, where red bars (observed) exceed blue bars (modeled), and in the scatterplot, where several points fall below the reference line. Despite this, the model effectively captures the temporal distribution of rainy days, with minimal false alarms or missed events. Overall, MSWEP performs very well at the Delia station, accurately reproducing daily precipitation patterns and intensity distributions, making it a reliable tool for hydrometeorological applications in this part of Sicily.

In contrast, the performance of MSWEP at the Linguaglossa station (Fig. 5), identified as one of the lowest-performing stations in this study, is somewhat weaker, with $r = 0.70$. While this still indicates a moderately strong relationship, the scatterplot reveals a wider spread of points and more frequent underestimations, particularly for extreme precipitation events exceeding 100 mm. The time series further highlights MSWEP's tendency to underestimate high-intensity precipitation, especially during extreme events between 2015 and 2020, where observed values surpassed 200 mm while modeled values remained significantly lower. Notably, while Delia experiences peak events below 100 mm, Linguaglossa records much more intense episodes, with rainfall sometimes exceeding 300 mm. This considerable difference in peak magnitudes amplifies the challenges for MSWEP, as the model appears increasingly limited in accurately reproducing such extreme values in high-precipitation zones. Although MSWEP captures the overall precipitation pattern and seasonal distribution at Linguaglossa, its ability to reflect the true magnitude of extreme events remains constrained. This performance gap is likely linked to the complex microclimatic conditions surrounding Linguaglossa, located on the slopes of Mount Etna, a region characterized by localized convective systems. These features not only lead to highly variable precipitation patterns but also generate extreme rainfall intensities that are difficult for satellite-based or blended products to resolve effectively.

While MSWEP performs well at both stations, its accuracy is notably higher in climatologically stable locations such as Delia. In

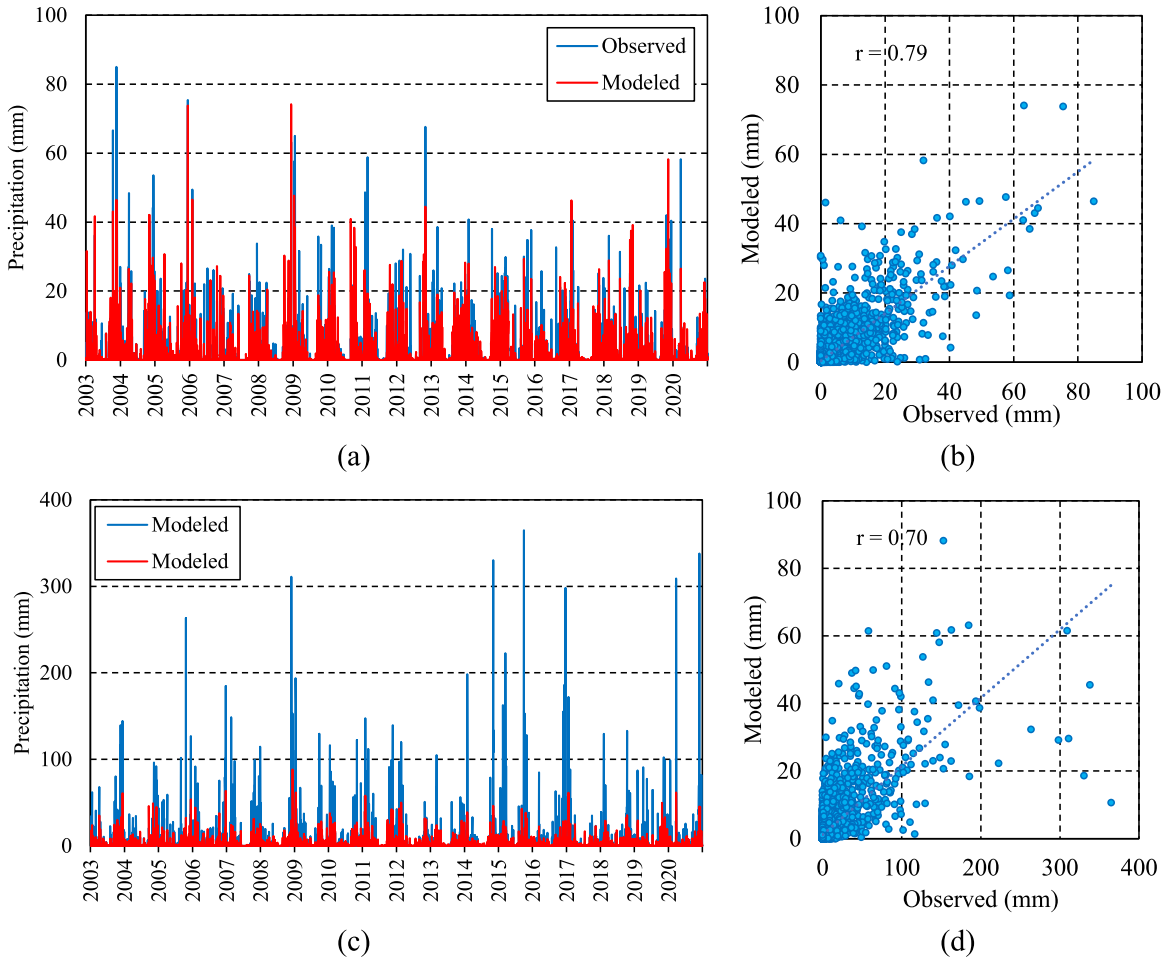


**Fig. 5.** Time series and scatter plots of precipitation at stations Delia (a and b) and Linguaglossa (c and d), comparing observed data with MSWEP data.

contrast, its limitations become more apparent in mountainous areas like Linguaglossa, particularly in capturing the full scale of extreme events. This comparison underscores the importance of local calibration or bias correction when applying gridded products in challenging environments, especially where extreme precipitation events play a critical hydrological role.

The GEV and POT analyses conducted at nine meteorological stations—Agrigento Mandrascava, Acate, Riesi, Delia, and Agrigento Scibica (shown in Fig. 6), and Caronia Buzza, Linguaglossa, Lentini, and San Fratello (shown in Fig. 6-continue)—provide a detailed and spatially differentiated assessment of the MSWEP dataset's ability to reproduce extreme rainfall events at the local scale. The stations were selected to reflect both the best- and worst-performing cases of MSWEP, ensuring that the analysis captures the full range of model behavior under varying hydroclimatic conditions. These two complementary methods offer a robust statistical framework for extreme value analysis, as they characterize the properties of events exceeding both annual maxima (GEV) and threshold values (POT), allowing a comprehensive evaluation of MSWEP's strengths and limitations in reproducing rare versus moderate extremes. This dual approach uniquely highlights how the model responds to different types of precipitation regimes, from frequent moderate rainfall to rare, intense convective events, across stations showing varying degrees of agreement between observed and modeled extremes.

At Agrigento Mandrascava, the empirical GEV density exhibits a distinct bimodal structure, with a primary mode centered around 35 mm (density $\approx 0.022$ mm$^{-1}$) and a secondary peak at 60–65 mm (density $\approx 0.022$ mm$^{-1}$), reflecting the coexistence of two distinct classes of extreme rainfall events: frequent moderate storms and less frequent, high-intensity events likely driven by mesoscale convection. In contrast, the MSWEP-based GEV collapses this complexity into a single narrow peak at 40 mm, reaching a maximum density of 0.05 mm$^{-1}$ —more than twice the height of the empirical peak—which indicates an over-concentration on moderate extremes at the expense of the heavy tail. Although MSWEP achieves lower information-criterion values (AIC = 130.34, BIC = 132.84) compared to the observational fit (AIC = 141.77, BIC = 144.27), the underrepresentation of the upper tail translates into systematic underestimation of return levels: for a 50-year return period, the observed and MSWEP estimates are 77 mm and 72 mm, respectively, while for the 100-year return period these rise to 81 mm versus 74 mm.

At Acate, the observed GEV density is characterized by an extended shoulder between 50 and 60 mm (density $\approx 0.025$ mm$^{-1}$) and a

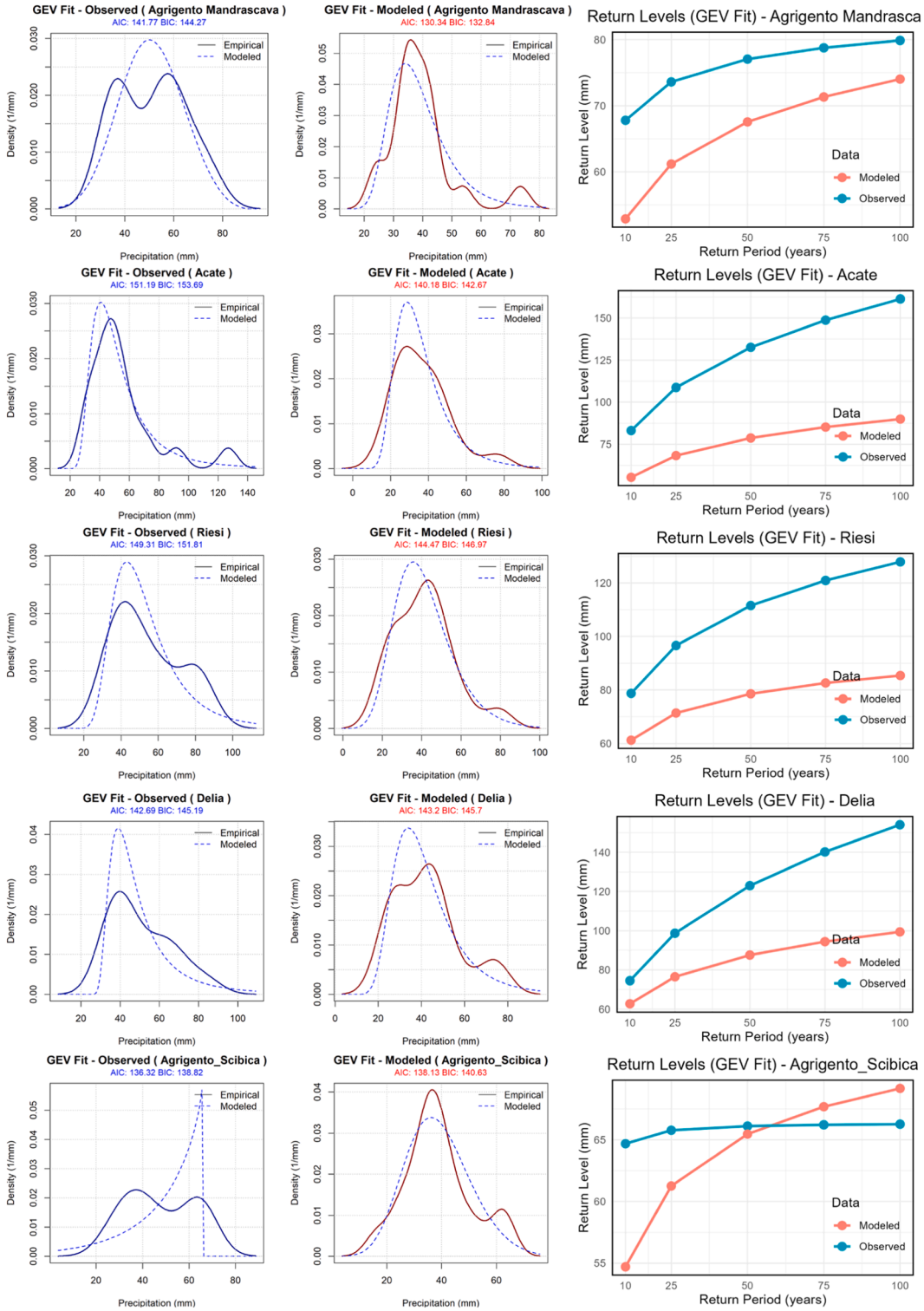


**Fig. 6.** Comparison of observed and modeled (MSWEP) annual maximum daily precipitation at stations (Agrigento Mandrascava, Acate, Riesi, Delia, and Agrigento Scibica). Left panels show GEV fits for observed data, middle panels show GEV fits for modeled (MSWEP) data, and right panels illustrate return levels for various return periods (10–100 years). AIC and BIC values are included to assess model fit quality.

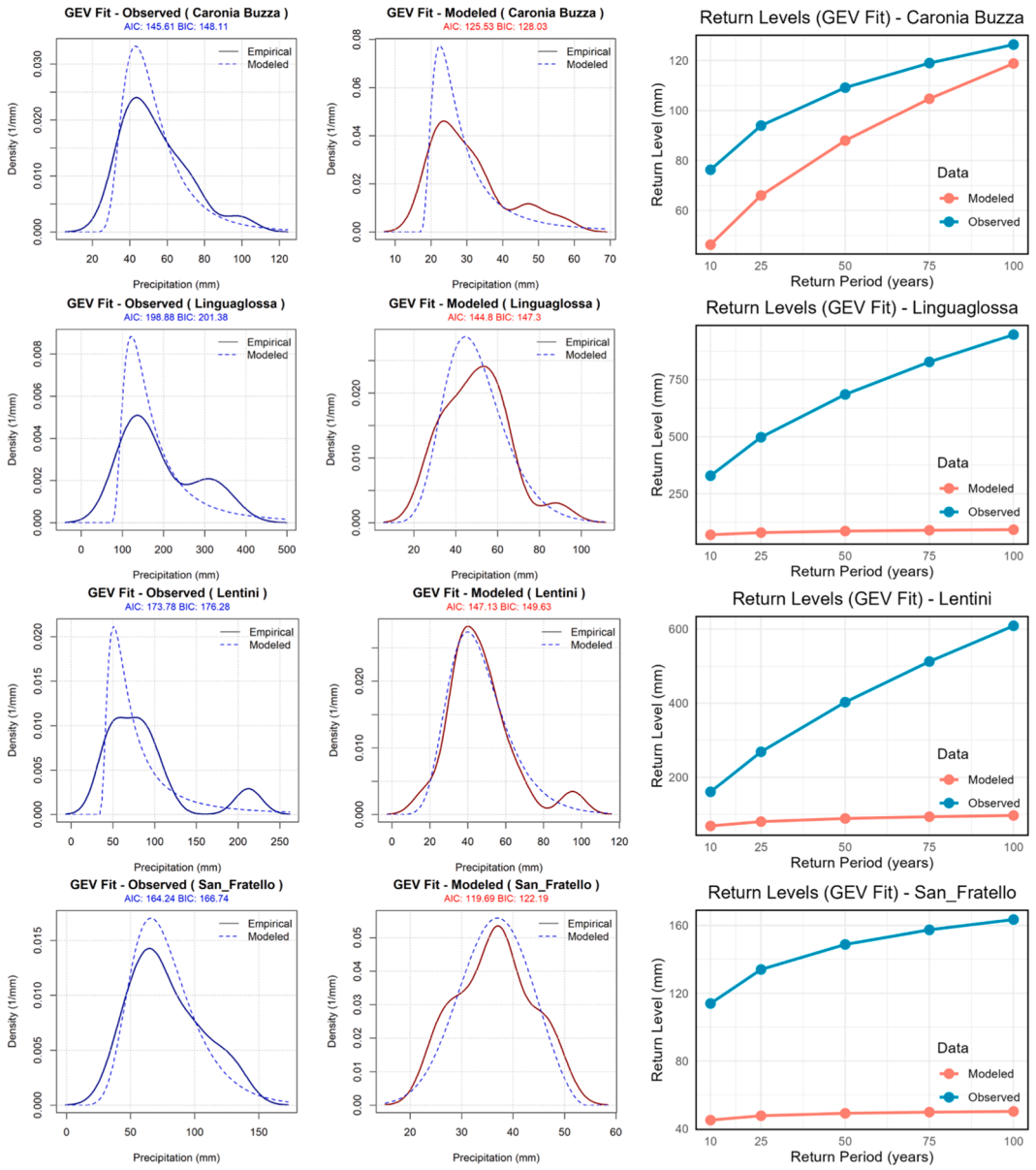


**Fig. 6.** (*continued*).

heavy right tail extending beyond 100 mm, indicative of substantial probability mass in rare, high-intensity events. By comparison, MSWEP is confined to the 30–35 mm range with a maximum density below 0.03 $mm^{-1}$ , revealing an inability to capture the empirical tail heaviness. Although the MSWEP fit improves information-criterion scores (AIC = 140.81, BIC = 146.27) relative to observations (AIC = 151.19, BIC = 153.62), it does so at the cost of severely underestimating extreme quantiles: the observed 50-year return level of 150 mm is reduced to 85 mm in MSWEP, while the 100-year return level drops from 160 mm to 93 mm.

At Riesi, empirical data show two subordinate modes at 40 mm and ~80 mm with a pronounced heavy tail, reflecting a mixture of frequent moderate events and rare, intense storms. MSWEP simplifies this structure into a single peak at 45 mm, significantly understating extreme events and underestimating the 100-year return level by approximately 40 mm (observed: 127 mm, modeled: 87 mm).

At Delia, the GEV fit suggests that MSWEP reasonably reproduces the general pattern of extreme precipitation, with broadly similar distribution shapes between empirical and modeled data. However, the observed density exhibits heavier upper tails, indicating higher variability and more extreme events than captured by the model. The empirical distribution peaks around 35 mm and has a longer right tail, whereas MSWEP shifts the peak to 45 mm and shortens the tail. While AIC/BIC values are comparable (observed: 142.69/145.19; MSWEP: 143.20/145.70), the 100-year return level is notably underestimated by approximately 50 mm (observed 150 mm vs. modeled 100 mm), and the 50-year level by 37 mm (observed 125 mm vs. modeled 88 mm).

At Agrigento Scibica, datasets show closer agreement. The empirical density displays a primary mode at 40 mm (≈0.025 $mm^{-1}$) and a secondary bump at 60 mm (≈0.02 $mm^{-1}$), while MSWEP smooths these into a broad mode centered at 35–40 mm with a maximum density of 0.04 $mm^{-1}$ , retaining tail behavior largely consistent with observations. Here, AIC/BIC values increase slightly under MSWEP (138.13/140.63) compared to observations (136.32/138.92), and 100-year return levels differ by only 3 mm (66 mm observed vs. 69 mm MSWEP).

At Caronia Buzza, the empirical GEV density exhibits a single, broad mode around 50 mm (intensity ≈ 0.03 $mm^{-1}$) with a modest shoulder extending toward 80 mm, reflecting the region's frequent moderate storms occasionally punctuated by more intense rainfall, likely influenced by local conditions and proximity to convective storm systems. In contrast, MSWEP reduces this complexity to a sharp peak at 25–30 mm (maximum intensity ≈ 0.08 $mm^{-1}$) with a truncated right tail, indicating that the model over-concentrates on moderate rainfall and fails to capture rarer, higher-intensity events. Although the AIC/BIC values decrease from 145.61/148.11 (observed) to 125.53/128.03 (MSWEP), reflecting an improved global fit, the tail representation is heavily compromised. Consequently, MSWEP underestimates return levels, with the 10-year return reduced from 110 mm to 20 mm and the 100-year from 130 mm to 10 mm, highlighting the dataset's limitation in representing both moderate and extreme rainfall in this transitional zone.

Linguaglossa shows a strongly heavy-tailed empirical density, with a primary mode near 120 mm (density ≈ 0.008 $mm^{-1}$) and a long right-hand tail extending up to 500 mm, indicative of extreme convective events intensified by mountainous terrain. MSWEP, however, is narrowly centered at 40–50 mm (peak density ≈ 0.028 $mm^{-1}$) with a severely curtailed tail, reflecting a marked inability to reproduce high-magnitude extremes. While AIC/BIC values drop from 198.88/201.38–144.80/147.30, suggesting a seemingly better fit, this comes at the expense of extreme quantiles, with the 10-year return level underestimated (~70 mm vs. 350 mm observed) and the 100-year level dramatically underestimated (~90 mm vs. 1000 mm observed). This underscores the challenge of modeling

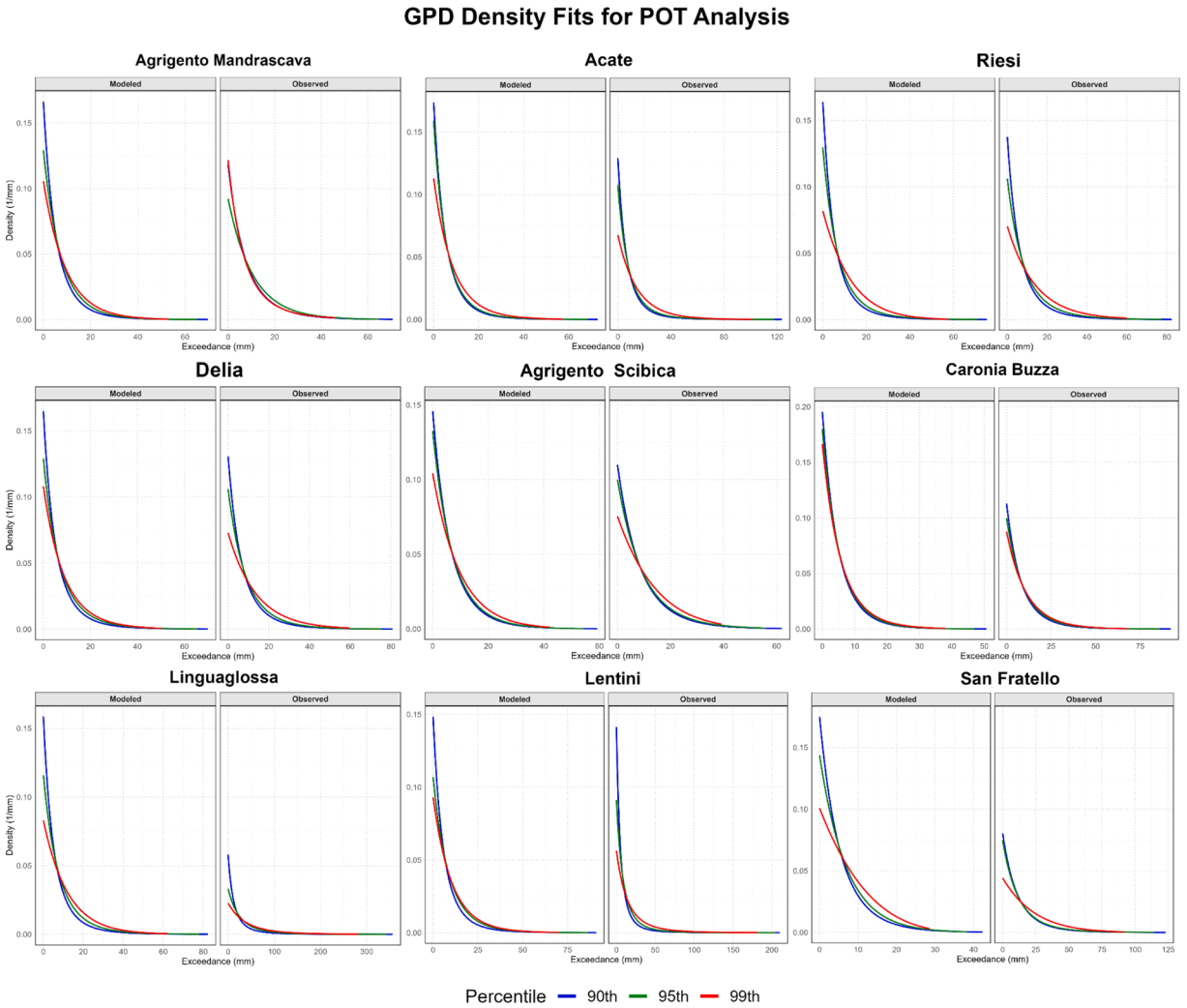


**Fig. 7.** GPD density fits for Peak Over Threshold (POT) analysis at multiple stations. Each panel compares modeled (left) and observed (right) precipitation exceedance distributions over selected thresholds (90th, 95th, and 99th percentiles). The plots illustrate how well the GPD fits the empirical exceedance data, with variations across stations and percentiles.

convectively-driven rainfall in northern Sicily.

At Lentini, the empirical density is bimodal, with a minor peak at ~60 mm (density ≈ 0.01 $mm^{-1}$) and a dominant peak around 100–150 mm, followed by a heavy tail extending to 250 mm, reflecting both frequent moderate events and rare intense convective or frontal rainfall events. MSWEP simplifies this structure into a single symmetric mode at 40–50 mm (density ≈ 0.022 $mm^{-1}$) with a sharply diminished tail, again underestimating the frequency and intensity of extreme events. Despite improved AIC/BIC values (from 173.78/176.28–147.13/149.63), return levels are dramatically underestimated: the 10-year level falls from 160 mm to 40 mm, and the 100-year level from 600 mm to 110 mm, highlighting the model's limitations in capturing the full complexity of bimodal extreme distributions.

San Fratello presents an empirical GEV density with a single, well-defined peak at 80 mm (density ≈ 0.013 $mm^{-1}$) and a moderate right tail extending beyond 150 mm, reflecting a mixture of regular moderate rainfall and less frequent high-intensity events. MSWEP centers at 35–45 mm with a peak density near 0.05 $mm^{-1}$ and a markedly shorter tail, resulting in a substantial underestimation of return levels: 10-year events are ~70 mm lower than observed, 50-year ~95 mm lower, and 100-year ~95 mm lower. Information criteria also drop considerably (from 164.24/168.74–119.69/122.19), suggesting an improved overall fit but masking the model's failure to represent extreme values, which are crucial for flood risk assessment.

Fig. 6. continued Comparison of observed and modeled (MSWEP) annual maximum daily precipitation at stations (Caronia Buzza, Linguaglossa, Lentini, and San Fratello). Left panels show GEV fits for observed data, middle panels show GEV fits for modeled (MSWEP) data, and right panels illustrate return levels for various return periods (10–100 years). AIC and BIC values are included to assess model fit quality.

An analysis of POT behavior across the stations reveals a systematic underrepresentation of extreme overflow events in MSWEP compared to observations, emphasizing the dataset's limitations in capturing the full spectrum of rare precipitation events. Fig. 7 illustrates the GPD exceedance curves for both observed and MSWEP-based extremes at each location, highlighting the divergence in the heavy-tail behavior.

At Agrigento Mandrascava, Riesi, and Delia, the observed %99 thresholds extend to 40–60 mm, whereas MSWEP remains constrained to 20–30 mm. This pronounced suppression of the upper tail suggests that MSWEP systematically underestimates the intensity of rare, high-impact events, which could be associated with mesoscale convective activity or localized topographic enhancement. Similarly, Acate, Caronia Buzza, and San Fratello exhibit notable discrepancies across both moderate and extreme precipitation ranges. While observed exceedances reach 75–100 mm at the %99 threshold, MSWEP rapidly decays around 30–40 mm, demonstrating the model's inability to represent convective storm-driven extremes and the underestimation of mid-range overflow events, which are critical for hydrological hazard assessment.

The mismatch becomes even more dramatic at Linguaglossa and Lentini, where the observed %99 tails extend to 200–300 mm, reflecting extreme convectively enhanced rainfall in mountainous terrain. MSWEP, by contrast, remains confined to 80–100 mm, severely truncating the distribution and failing to reproduce the extreme upper tail. Only Agrigento Scibica exhibits relatively smaller deviations; MSWEP captures a comparable range (8–40 mm) to the observed 10–45 mm but still slightly underestimates the upper thresholds, particularly beyond the 95th percentile. Even in this best-performing case, the subtle discrepancies in tail behavior highlight the limits of gridded datasets for reproducing localized extremes.

These limitations carry substantial implications for applications that rely on accurate high-quantile estimation, such as hydrological design, flood risk mapping, and early warning systems. Although MSWEP remains the best-performing dataset in broader intercomparisons, its consistent underrepresentation of upper-tail extremes underscores the need for methodological refinements when applied in hazard-prone or topographically complex regions. Potential strategies include station-specific calibration, bias correction methods, and integration with high-resolution ground-based observations, which are particularly important in Mediterranean environments like Sicily, where the interaction between rugged terrain and dynamic atmospheric processes generates highly localized and intense hydroclimatic extremes.

## 4. Discussion

### *4.1. Contextualizing dataset performance: insights from the Mediterranean region and beyond*

The evaluation of 11 global precipitation datasets over Sicily reveals critical insights into their performance across diverse climatic conditions. MSWEP, ERA5-Land, and HydroGFD emerged as the top performers due to their multi-source blending strategies, which integrate satellite retrievals, reanalysis outputs, and gauge observations. This integration enhances their ability to resolve spatial precipitation patterns, particularly in southern Sicily. However, their superiority declines in mountainous areas such as Mount Etna, where localized convective systems generates extreme rainfall. The systematic underestimation of high-intensity events by these datasets highlights a fundamental limitation: global products often smooth out small-scale atmospheric processes due to coarse spatial resolutions or inadequate representation of microclimatic dynamics.

This variability in performance is further reflected in the pronounced spatial differences in dataset accuracy across the island. For example, higher correlations ($r = 0.79$) are observed in Delia, Southern Sicily, while weaker performance ($r = 0.70$) is noted in highly variable microclimates areas such as Linguaglossa (close to the Etna, Northeastern Sicily). These discrepancies align with Sicily's climatic dichotomy, where southern regions, characterized by stable, low-intensity rainfall, are well-captured by satellite and reanalysis products. In contrast, northern mountainous zones, where precipitation extremes are amplified by elevation and proximity to moisture-laden air masses, present more challenges. This highlights the importance of considering microclimates when evaluating dataset performance.

In addressing the study's first research question, the findings confirm that blended products like MSWEP are optimal for daily precipitation estimation but require caution in extreme scenarios. Further, the extreme value analysis, focusing on high-intensity rainfall events, answers the second research question by revealing critical gaps in dataset accuracy. While MSWEP reliably captures moderate extremes (e.g., 50-year return levels), it underestimates severe events (e.g., 300 mm peaks at Linguaglossa). This discrepancy underscores a broader issue: global datasets prioritize broad-scale consistency over localized extremes, which are often driven by short-duration, high-intensity processes that current sensor technologies or model parameterizations struggle to resolve.

Overall, the assessment of these global precipitation datasets in Sicily not only reveals performance differences across products but also highlights the challenges of capturing precipitation patterns in regions with varying climatic influences. To place these findings within a broader geographical and methodological framework, comparisons were drawn with previous evaluations of precipitation datasets across the Mediterranean basin.

The performance of global precipitation datasets evaluated in this study for the region of Sicily reveals both similarities and contextual differences when compared with previous studies conducted across various parts of the Mediterranean Basin. A common challenge consistently highlighted in all these studies is the difficulty of accurately representing precipitation at the regional scale due to microclimatic variability. This issue becomes especially pronounced in areas where convective systems strongly influence rainfall dynamics.

Temporal and spatial resolution differences emerge as critical factors affecting dataset performance evaluations. For instance, Hatzianastassiou et al. (2016) assessed monthly precipitation across the Mediterranean Basin for the period 1979–2010 using GPCPv2 satellite-based data. While they reported a high annual correlation with rain gauge observations ($R = 0.78$), the coarse spatial resolution ($2.5^\circ \times 2.5^\circ$) and monthly temporal scale limited the dataset's ability to accurately capture extreme rainfall events. In contrast, in our study, based on daily-scale assessments and higher-resolution comparisons, GPCPv2 demonstrated considerably weaker performance, ranking near the bottom of the evaluated datasets. This discrepancy underscores how both temporal resolution and analysis scale can significantly impact the perceived skill of precipitation datasets.

Differences between model-based and satellite-based precipitation products offer a valuable perspective for performance evaluation. Mascaro et al. (2018), in their assessment of EURO-CORDEX regional climate models at spatial resolutions of $0.11^\circ$ and $0.44^\circ$ over Sardinia, found that most models successfully captured the typical Mediterranean pattern of increasing precipitation with elevation. Importantly, they also documented substantial spatial biases, with annual precipitation deviations reaching up to $\pm 60$ %, underscoring the challenges these models face in accurately representing precipitation in highly variable microclimates. These findings strongly suggest that model-based products tend to oversmooth precipitation fields in mountainous regions, leading to an underrepresentation of localized high-intensity events.

Our analysis revealed similar tendencies: coarse-resolution datasets showed limited sensitivity to spatial gradients and consistently underestimated extreme precipitation events. However, unlike Mascaro et al. (2018), we found no significant correlation between dataset performance metrics and altitude, indicating that elevation alone does not fully explain spatial variability in these products.

Geographical proximity further highlights the value of cross-validation across regions. Caroletti et al. (2019) focused on Calabria, a region climatically comparable to Sicily, analyzing CHIRPS, E-OBS, and several GCM-RCM combinations at a monthly scale. E-OBS showed high correlation with ground observations ($r = 0.97$), yet performed moderately in terms of relative errors. In our study, which employed daily data, E-OBS ranked fifth among the eleven evaluated datasets, demonstrating acceptable performance but lagging behind higher-resolution datasets such as MSWEP and ERA5-Land. This contrast suggests that both temporal scale and sensitivity to extremes can influence the comparative performance of a dataset.

The study by Akbas and Ozdemir (2024), which evaluated daily precipitation datasets for hydrological modeling in mountainous catchments in Türkiye, provides another relevant benchmark due to its methodological similarity to the present study. Their analysis identified APHRODITE ($0.25^\circ$ resolution) and CPC ($0.5^\circ$ resolution) as the best-performing products. However, due to the unavailability of APHRODITE for Sicily and the coarse spatial resolution of CPC, these datasets were not included in our evaluation. Notably, the datasets shared between the two studies, ERA5-Land, MSWEP, and HydroGFD, were consistently among the top performers. This finding reinforces the ability of these products to deliver reliable precipitation estimates even under complex conditions. Furthermore, both studies identified the PERSIANN family and TRMM as underperforming datasets, reaffirming their limitations in capturing the magnitude and frequency of precipitation events in regions with highly variable microclimates.

The comparative analysis of global precipitation datasets across Mediterranean regions reaffirms a persistent challenge: the accurate representation of rainfall, such as those found in Sicily. Products with finer spatial resolution, most notably ERA5-Land and MSWEP, consistently outperformed coarser datasets, particularly in resolving spatial gradients and capturing localized precipitation dynamics. These findings are consistent with previous studies conducted in similarly heterogeneous Mediterranean settings and underscore the added value of high-resolution datasets when modeling precipitation in environments shaped by diverse climatic influences.

### 4.2. *Precipitation datasets and their error sources*

Each category of global precipitation dataset, satellite-based, reanalysis, and blended, has distinct sources of error that affect the accuracy of daily and extreme rainfall estimates. Understanding these uncertainties is essential to interpret performance differences observed across Sicily.

Satellite-based datasets rely on sensors measuring radiances in the microwave (PMW) or infrared (IR) spectrum. IR-based algorithms typically use cloud-top temperature thresholds to infer rainfall intensity, assuming that colder cloud tops correspond to deeper convective systems. However, this approach often fails to detect rainfall from shallow or warm clouds, leading to systematic

underestimation, particularly in Mediterranean environments where stratiform precipitations are frequent (Peinó et al. 2024). Passive microwave retrievals improve detection accuracy but remain limited by coarse spatial resolution and infrequent satellite overpasses. Moreover, retrieval algorithms are sensitive to surface emissivity, introducing additional uncertainty over coastal and mountainous regions such as northeastern Sicily (Katsanos et al., 2024). As a result, satellite precipitation products such as GPM IMERG or PERSIANN tend to underestimate heavy convective rainfall and struggle to represent the full intensity range of localized extremes common in the Mediterranean.

Reanalysis datasets, such as ERA5-Land and GLDAS-2, generate precipitation fields through numerical weather prediction models constrained by data assimilation. Their accuracy depends on the quality of assimilated observations and on the parameterization of convective processes. Systematic biases arise from simplified representations of convection, cloud microphysics, and terrain interactions, which tend to smooth high-intensity precipitation peaks (Cavalleri et al., 2024). These limitations are consistent with our finding that reanalysis products underestimate extreme precipitation along the slopes of Mount Etna, where complex topography and convective uplift amplify rainfall variability.

Blended products such as MSWEP and HydroGFD combine satellite retrievals, reanalysis outputs, and gauge observations to exploit the strengths of each source. While this fusion generally enhances accuracy and spatial consistency, residual errors can arise from differences in the temporal and spatial resolutions of the input data. Mismatches in scaling or bias-correction procedures between the component datasets may propagate into the final merged product, producing localized inconsistencies (Bartsotas et al., 2018). Nonetheless, the blending approach mitigates single-source biases, explaining the superior performance of MSWEP and HydroGFD in reproducing daily precipitation across Sicily.

Overall, the spatial patterns observed in dataset performance reflect these inherent error structures. The underestimation of intense rainfall in mountainous and convective regions mirrors well-known retrieval and modeling limitations, emphasizing the need for local calibration and integration of high-resolution ground observations to enhance dataset reliability in hydro-climatically complex Mediterranean environments.

### *4.3. Practical implications, limitations and future research directions*

The findings offer valuable guidance for stakeholders involved in hydrology and disaster management, particularly regarding the selection and use of global precipitation datasets. For short-term hydrological applications in Sicily, MSWEP and ERA5-Land emerge as the most reliable datasets due to their balanced accuracy in capturing daily precipitation patterns. However, in flood-prone mountainous regions, such as those around Mount Etna, it is crucial to supplement these datasets with local gauge data or apply bias correction techniques to improve the representation of extreme events. This ensures that daily hydrological models can more accurately simulate flood risks in areas prone to intense rainfall.

In disaster risk management, global datasets used in flood forecasting systems should incorporate local calibration, statistical adjustments, or multi-dataset merging to address the underestimation of daily or short-term extremes. Such adjustments are important for designing infrastructure, such as drainage systems, that can withstand high-intensity daily rainfall events. Without these refinements, the risk of infrastructure failure during extreme rainfall events remains high.

For operational monitoring, investing in denser ground-based station networks, particularly in northern Sicily, can improve real-time data availability. Agencies like SIAS can support rapid recalibration of global datasets, ensuring their continued reliability for short-term forecasting and flood management.

While the study provides a comprehensive evaluation of global precipitation datasets over Sicily, several limitations should be noted. First, the analysis focused exclusively on daily precipitation, overlooking sub-daily or hourly extremes, which are particularly relevant for flash flood forecasting. Second, while a dense network of monitoring stations was used, data gaps or localized measurement errors may have introduced uncertainty into the validation process. Third, the temporal scope was limited to the period 2003–2023 to ensure consistency across datasets, which constrains assessment of longer-term dataset stability.

Additionally, the study concentrated on statistical accuracy at the daily scale without evaluating hydrological performance when the datasets are integrated into models (e.g., runoff or soil moisture simulations). The extreme precipitation analysis was also limited to a selection of stations and events, potentially underrepresenting spatial variability across microclimatic zones, particularly in regions like Mount Etna.

Future research should consider higher temporal resolution datasets to evaluate their suitability for short-term forecasting and early warning applications. Expanding the analysis to include hydrological modeling would provide insights into the operational relevance of these datasets. Moreover, developing bias correction or dataset blending approaches tailored to Sicily's unique hydroclimatic conditions could improve accuracy for extreme rainfall events. Finally, continuous validation with newly observed extreme events will be essential to ensure that global precipitation products remain reliable tools for short-term hydrological applications and flood risk management.

## 5. Conclusion

This study systematically evaluated 11 global precipitation datasets, including satellite-based, reanalysis, and blended products, against ground-based observations across Sicily. The major findings can be summarized as follows:

- Dataset performance. MSWEP consistently provided the most accurate daily precipitation estimates, followed by ERA5-Land and HydroGFD.

- Precipitation patterns. All evaluated datasets performed well in capturing overall precipitation patterns but struggled with high-intensity events, particularly in mountainous regions such as the northeastern slopes of Mount Etna.
- Extreme events. Analyses using the GEV distribution and POT method showed that MSWEP reliably represents moderate extremes but underestimates the magnitude and frequency of severe precipitation events.
- Implications for practice. Selection of precipitation datasets for hydrological modeling and disaster risk management should consider dataset limitations, and the integration of high-resolution local observations is recommended to improve extreme event representation.

Overall, the study highlights the strengths and limitations of current global precipitation products in regions with microclimatic variability. Continued methodological improvements and local calibration are essential to enhance dataset reliability and better support water resource management and flood mitigation in Sicily and similar Mediterranean regions.

## CRediT authorship contribution statement

**DI NUNNO Fabio:** Writing – original draft, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Giovanni de Marinis:** Writing – original draft, Visualization, Supervision, Investigation. **Mehmet Berkant Yıldız:** Writing – original draft, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Francesco Granata:** Writing – original draft, Visualization, Validation, Supervision, Methodology, Investigation, Data curation, Conceptualization.

## Declaration of Competing Interest

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

## Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.ejrh.2025.103062.

## Data availability

Data Info are reported in Data Availability statement

## References

Acero, F.J., García, J.A., Gallego, M.C., 2011. Peaks-over-threshold study of trends in extreme rainfall over the Iberian Peninsula. J. Clim. 24, 1089–1105. https://doi.org/10.1175/2010JCLI3627.1.

Agilan, V., Umamahesh, N.V., Mujumdar, P.P., 2021. Influence of threshold selection in modeling peaks over threshold based nonstationary extreme rainfall series. J. Hydrol. 593, 125625. https://doi.org/10.1016/j.jhydrol.2020.125625.

Akbas, A., Ozdemir, H., 2024. Comparing satellite, reanalysis, fused and gridded (in situ) precipitation products over Türkiye. Int J. Clim. 44, 5873–5889. https://doi.org/10.1002/joc.8671.

Ashouri, H., Hsu, K.-L., Sorooshian, S., et al., 2015. PERSIANN-CDR: daily precipitation climate data record from multisatellite observations for hydrological and climate studies. Bull. Am. Meteor. Soc. 96, 69–83. https://doi.org/10.1175/BAMS-D-13-00068.1.

Bartsotas, N.S., Anagnostou, E.N., Nikolopoulos, E.I., Kallos, G., 2018. Investigating satellite precipitation uncertainty over complex terrain. J. Geophys. Res. Atmos. 123, 5346–5359. https://doi.org/10.1029/2017JD027559.

Beck, H.E., Vergopolan, N., Pan, M., et al., 2017. Global-scale evaluation of 22 precipitation datasets using gauge observations and hydrological modeling. Hydrol. Earth Syst. Sci. 21, 6201–6217. https://doi.org/10.5194/hess-21-6201-2017.

Beck, H.E., Wood, E.F., Pan, M., et al., 2019. MSWEP V2 global 3-hourly 0.1° precipitation: methodology and quantitative assessment. Bull. Am. Meteorol. Soc. 100, 473–500. https://doi.org/10.1175/BAMS-D-17-0138.1.

Berg, P., Almén, F., Bozhinova, D., 2021. HydroGFD3.0 (Hydrological global forcing data): a 25 km global precipitation and temperature data set updated in near-real time. Earth Syst. Sci. Data 13, 1531–1545. https://doi.org/10.5194/essd-13-1531-2021.

Berg, P., Donnelly, C., Gustafsson, D., 2018. Near-real-time adjusted reanalysis forcing data for hydrology. Hydrol. Earth Syst. Sci. 22, 989–1000.

Caracciolo, D., Francipane, A., Viola, F., et al., 2018. Performances of GPM satellite precipitation over the two major Mediterranean islands. Atmos. Res. 213, 309–322. https://doi.org/10.1016/j.atmosres.2018.06.010.

Caroletti, G.N., Coscarelli, R., Caloiero, T., 2019. Validation of satellite, reanalysis and RCM data of monthly rainfall in Calabria (Southern Italy). Remote Sens. 11 (13), 1625. https://doi.org/10.3390/rs11131625.

Cauteruccio, A., Chinchella, E., Lanza, L.G., 2024. The overall collection efficiency of catching-type precipitation gauges in windy conditions. Water Resour. Res. 60, e2023WR035098. https://doi.org/10.1029/2023WR035098.

Cavalleri, F., Lussana, C., Viterbo, F., et al., 2024. Multi-scale assessment of high-resolution reanalysis precipitation fields over Italy. Atmos. Res. 312, 107734. https://doi.org/10.1016/j.atmosres.2024.107734.

Cheng, Y., Zhang, X., Wang, K., et al., 2025. Multidimensional evaluation of satellite-based and reanalysis-based precipitation datasets in the Tibetan Plateau. J. Hydrol. 660, 133364. https://doi.org/10.1016/j.jhydrol.2025.133364.

Cornes, R.C., Van Der Schrier, G., Van Den Besselaar, E.J.M., Jones, P.D., 2018. An ensemble version of the E-OBS temperature and precipitation data sets. J. Geophys. Res. Atmos. 123, 9391–9409.

Dallan, E., Bagarello, V., Ferro, V., et al., 2025. Evaluation and projected changes in rainfall erosivity: Topography dependence revealed by convection-permitting climate projections for the Mediterranean island of Sicily. CATENA 254, 108975. https://doi.org/10.1016/j.catena.2025.108975.

Davolio, S., Della Fera, S., Laviola, S., Miglietta, M.M., Levizzani, V., 2020. Heavy Precipitation over Italy from the Mediterranean Storm "Vaia" in October 2018: assessing the role of an atmospheric river. Mon. Weather Rev. 148, 3571–3588. https://doi.org/10.1175/MWR-D-20-0021.1.
Dinku, T., Chidzambwa, S., Ceccato, P., Connor, S.J., Ropelewski, C.F., 2008. Validation of high-resolution satellite rainfall products over complex terrain. Int. J. Remote Sens. 29 (14), 4097–4110. https://doi.org/10.1080/01431160701772526.
Du, Y., Wang, D., Zhu, J., et al., 2022. Intercomparison of multiple high-resolution precipitation products over China: climatology and extremes. Atmos. Res. 278, 106342. https://doi.org/10.1016/j.atmosres.2022.106342.
ECMWF 2017. Copernicus Climate Change Service (C3S): ERA5: Fifth generation of ECMWF atmospheric reanalyses of the global climate. Copernicus Climate Change Service Climate Data Store (CDS), date of access.
Flitcroft, I.D., Milford, J.R., Dugdale, G., 1989. Relating point to area average rainfall in semiarid West Africa and the implications for rainfall estimates derived from satellite data. J. Appl. Meteorol. Clim. 28, 252–266. https://doi.org/10.1175/1520-0450(1989)028<0252:RPTAAR>2.0.CO;2.
Gentilucci, M., Rossi, A., Pelagagge, N., Aringoli, D., Barbieri, M., Pambianchi, G., 2023. GEV analysis of extreme rainfall: comparing different time intervals to analyse model response in terms of return levels in the study area of central Italy. Sustainability 15 (15), 11656. https://doi.org/10.3390/su151511656.
Gupta, H.V., Kling, H., Yilmaz, K.K., Martinez, G.F., 2009. Decomposition of the mean squared error and NSE performance criteria: implications for improving hydrological modelling. J. Hydrol. 377, 80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003.
Hatzianastassiou, N., Papadimas, C.D., Lolis, C.J., Bartzokas, A., Levizzani, V., Pnevmatikos, J.D., Katsoulis, B.D., 2016. Spatial and temporal variability of precipitation over the Mediterranean Basin based on 32-year satellite global precipitation climatology project data, part I: evaluation and climatological patterns. Int. J. Clim. 36, 4741–4754. https://doi.org/10.1002/joc.4666.
Haylock, M.R., Hofstra, N., Klein Tank, A.M.G., et al., 2008. A European daily high-resolution gridded data set of surface temperature and precipitation for 1950–2006. J. Geophys. Res. Atmos. 113.
Hersbach, H., Bell, B., Berrisford, P., 2020. The ERA5 global reanalysis. Q J. R. Meteorol. Soc. 146, 1999–2049. https://doi.org/10.1002/qj.3803.
Hong, Y., Hsu, K.-L., Sorooshian, S., Gao, X., 2004. Precipitation estimation from remotely sensed imagery using an artificial neural network cloud classification system. J. Appl. Meteor. 43, 1834–1853. https://doi.org/10.1175/JAM2173.1.
Hou, A.Y., Kakar, R.K., Neeck, S., et al., 2014. The global precipitation measurement mission. Bull. Am. Meteorol. Soc. 95, 701–722. https://doi.org/10.1175/BAMS-D-13-00164.1.
Hsu, K., Gao, X., Sorooshian, S., Gupta, H.V., 1997. Precipitation estimation from remotely sensed information using artificial neural networks. J. Appl. Meteorol. 36, 1176–1190.
Huffman G.J., Bolvin D.T. (2018) TRMM and other data precipitation data set documentation. NASA, Greenbelt, USA 28:1.
Huffman, G.J., Bolvin, D.T., Nelkin, E.J., et al., 2007. The TRMM multisatellite precipitation analysis (TMPA): quasi-global, multiyear, combined-sensor precipitation estimates at fine scales. J. Hydrometeorol. 8, 38–55. https://doi.org/10.1175/JHM560.1.
Huffman G.J., Stocker E.F., Bolvin D.T., et al 2019. GPM IMERG final precipitation L3 half hourly 0.1 ° x 0.1 ° V06, Greenbelt, MD, Goddard Earth Sciences Data and Information Services Center (GES DISC). Goddard Earth Sci Data Inf Serv Cent (GES DISC).
Katsanos, D., Retalis, A., Kalogiros, J., Psiloglou, B.E., Roukounakis, N., Anagnostou, M., 2024. Performance evaluation of satellite precipitation products during extreme events—the case of the medicane Daniel in Thessaly, Greece. Remote Sens. 16 (22), 4216. https://doi.org/10.3390/rs16224216.
Keikhosravi-Kiany, M.S., Masoodian, S.A., Balling Jr, R.C., Darand, M., 2022. Evaluation of tropical rainfall measuring mission, integrated multi-satellite retrievals for GPM, climate hazards centre infrared precipitation with station data, and european centre for medium-range weather forecasts reanalysis v5 data in estimating precipitation and capturing meteorological droughts over Iran. Int. J. Climatol. 42 (4), 2039–2064. https://doi.org/10.1002/joc.7351.
Krause, P., Boyle, D.P., Bäse, F., 2005. Comparison of different efficiency criteria for hydrological model assessment. Adv. Geosci. 5, 89–97. https://doi.org/10.5194/adgeo-5-89-2005.
Leeper, R.D., Kochendorfer, J., 2015. Evaporation from weighing precipitation gauges: impacts on automated gauge measurements and quality assurance methods. Atmos. Meas. Tech. 8, 2291–2300. https://doi.org/10.5194/amt-8-2291-2015.
Lei, H., Li, H., Zhao, H., et al., 2021. Comprehensive evaluation of satellite and reanalysis precipitation products over the eastern Tibetan plateau characterized by a high diversity of topographies. Atmos. Res. 259, 105661. https://doi.org/10.1016/j.atmosres.2021.105661.
Li, B., Rodell, M., Sheffield, J., et al., 2019. Long-term, non-anthropogenic groundwater storage changes simulated by three global-scale hydrological models. Sci. Rep. 9, 10746. https://doi.org/10.1038/s41598-019-47219-z.
Lo Conti, F., Hsu, K., Noto, L., Sorooshian, S. 2012. Evaluating several satellite precipitation estimates and global ground-based dataset on Sicily (Italy). In: Proc. SPIE 8531, Remote Sensing for Agriculture, Ecosystems, and Hydrology XIV, 853101 (October 19, 2012) [10.1117/12.2014606].
Lu, S., Veldhuis, M. ten, van de Giesen, N., 2020. A methodology for multiobjective evaluation of precipitation products for extreme weather (in a data-scarce environment). J. Hydrometeorol. 21, 1223–1244. https://doi.org/10.1175/JHM-D-19-0157.1.
Mascaro, G., Viola, F., Deidda, R., 2018. Evaluation of precipitation from EURO-CORDEX regional climate simulations in a small-scale Mediterranean site. J. Geophys. Res. Atmos. 123, 1604–1625. https://doi.org/10.1002/2017JD027463.
Maugeri, M., Brunetti, M., Garzoglio, M., Simolo, C., 2015. High-resolution analysis of 1 day extreme precipitation in Sicily. Nat. Hazards Earth Syst. Sci. 15, 2347–2358. https://doi.org/10.5194/nhess-15-2347-2015.
Moriasi, D.N., Arnold, J.G., Van Liew, M.W., et al., 2007. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. Trans. ASABE 50, 885–900.
Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., et al., 2021. ERA5-land: a state-of-the-art global reanalysis dataset for land applications. Earth Syst. Sci. Data 13, 4349–4383. https://doi.org/10.5194/essd-13-4349-2021.
Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models part I — a discussion of principles. J. Hydrol. 10, 282–290. https://doi.org/10.1016/0022-1694(70)90255-6.
Nešpor, V., Sevruk, B., 1999. Estimation of wind-induced error of rainfall gauge measurements using a numerical simulation. J. Atmos. Ocean Technol. 16, 450–464. https://doi.org/10.1175/1520-0426(1999)016<0450:EOWIEO>2.0.CO;2.
Nguyen, P., Ombadi, M., Gorooh, V.A., et al., 2020a. PERSIANN dynamic infrared–rain rate (PDIR-Now): a near-real-time, quasi-global satellite precipitation dataset. J. Hydrometeorol. 21, 2893–2906. https://doi.org/10.1175/JHM-D-20-0177.1.
Nguyen, P., Shearer, E.J., Ombadi, M., et al., 2020b. PERSIANN dynamic infrared–rain rate model (PDIR) for high-resolution, real-time satellite precipitation estimation. Bull. Am. Meteorol. Soc. 101, E286–E302. https://doi.org/10.1175/BAMS-D-19-0118.1.
Nguyen, P., Shearer, E.J., Tran, H., et al., 2019. The CHRS Data Portal, an easily accessible public repository for PERSIANN global satellite precipitation data. Sci. data 6, 1–10.
Noto, L., Beikahmadi, N., Pumo, D., Francipane, A., 2022. An artificial intelligence–based blending of satellite products across Mediterranean Island of Sicily, Italy using GPM-IMERG V06 final run. EMS Annu. Meet. Abstr. 19. https://doi.org/10.5194/ems2022-503].
Peinó, E., Bech, J., Udina, M., Polls, F., 2024. Disentangling satellite precipitation estimate errors of heavy rainfall at the daily and sub-daily scales in the Western Mediterranean. Remote Sens. 16 (3), 457. https://doi.org/10.3390/rs16030457.
Polaris 2025. Evento alluvionale in Sicilia e Calabria. ⟨https://polaris.irpi.cnr.it/event/evento-alluvionale-in-sicilia-e-calabria/⟩. Accessed 19 Apr 2025.
Pollock, M.D., O'Donnell, G., Quinn, P., et al., 2018. Quantifying and mitigating wind-induced undercatch in rainfall measurements. Water Resour. Res 54, 3863–3875. https://doi.org/10.1029/2017WR022421.
Rao, P., Wang, F., Yuan, X., et al., 2024. Evaluation and comparison of 11 sets of gridded precipitation products over the Qinghai-Tibet Plateau. Atmos. Res. 302, 107315. https://doi.org/10.1016/j.atmosres.2024.107315.
Schellander, H., Lieb, A., Hell, T., 2019. Error structure of metastatistical and generalized extreme value distributions for modeling extreme rainfall in Austria. Earth Space Sci. 6, 1616–1632. https://doi.org/10.1029/2019EA000557.
Scicchitano, G., Scardino, G., Monaco, C., et al., 2021. Comparing impact effects of common storms and Medicanes along the coast of south-eastern Sicily. Mar. Geol. 439, 106556.

Skofronick-Jackson, G., Petersen, W.A., Berg, W., et al., 2017. The global precipitation measurement (GPM) mission for science and society. Bull. Am. Meteorol. Soc. 98, 1679–1695. https://doi.org/10.1175/BAMS-D-15-00306.1.
Sorooshian, S., Duan, Q., Gupta, V.K., 1993. Calibration of rainfall-runoff models: application of global optimization to the sacramento soil moisture accounting model. Water Resour. Res. 29, 1185–1194.
Wilks, D.S., 2011. Statistical Methods in the Atmospheric Sciences. Academic press.
Willmott, C.J., 1981. On the validation of models. Phys. Geogr. 2, 184–194. https://doi.org/10.1080/02723646.1981.10642213.
Zambrano-Bigiarini, M., 2014. Goodness-of-fit functions for comparison of simulated and observed hydrological time series. R Packag Version 03-8.