Saudi Rainfall (SaRa): Hourly 0.1° Gridded Rainfall (1979–Present) for Saudi Arabia via Machine Learning Fusion of Satellite and Model Data

Xuetong Wang[1], Raied S. Alharbi[2], Oscar M. Baez-Villanueva[3], Amy Green[4,5], Matthew F. McCabe[1], Yoshihide Wada[1], Albert I.J.M. Van Dijk[6], Muhammad A. Abid[7,8], and Hylke E. Beck[1,9]

[1]King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia
[2]Department of Civil Engineering, College of Engineering, King Saud University, Riyadh, Saudi Arabia
[3]Hydro-Climate Extremes Lab (H-CEL), Ghent University, Ghent, Belgium
[4]School of Engineering, Newcastle University, Newcastle upon Tyne, UK
[5]Tyndall Centre for Climate Change Research, Newcastle University, Newcastle upon Tyne, UK
[6]Fenner School of Environment & Society, Australian National University, Canberra, ACT, Australia
[7]Atmospheric, Oceanic and Planetary Physics (AOPP), Department of Physics, University of Oxford, UK
[8]National Centre for Atmospheric Science (NCAS), UK
[9]GloH2O LLC, Princeton, USA

Correspondence: Hylke E. Beck (hylke.beck@gloh2o.org)

Abstract. We introduce Saudi Rainfall (SaRa), a gridded historical and near real-time precipitation (P) product specifically designed for the Arabian Peninsula, one of the most arid, water-stressed, and data-sparse regions on Earth. The product has an hourly 0.1° resolution spanning from 1979 to the present and is continuously updated with a latency of less than two hours. The algorithm underpinning the product involves 18 machine learning model stacks trained for different combinations of satellite and (re)analysis P products along with several static predictors. As a training target, hourly and daily P observations from gauges in Saudi Arabia (n=113) and globally (n=14,256) are used. To evaluate the performance of SaRa, we carried out the most comprehensive evaluation of gridded P products in the region to date, using observations from independent gauges (randomly excluded from training) in Saudi Arabia as a reference (n=119). Among the 20 evaluated P products, our new product, SaRa, consistently ranked first across all evaluation metrics, including the Kling-Gupta Efficiency (KGE), correlation, bias, peak bias, wet days bias, and critical success index. Notably, SaRa achieved a median KGE — a summary statistic combining correlation, bias, and variability — of 0.36, while widely used non-gauge-based products such as CHIRP, ERA5, GSMaP V8, and IMERG-L V07 achieved values of -0.07, 0.21, -0.13, and -0.39, respectively. SaRa also outperformed four gauge-based products such as CHIRPS V2, CPC Unified, IMERG-F V07, and MSWEP V2.8 which had median KGE values of 0.17, -0.03, 0.29, and 0.20, respectively. Our new P product — available at www.gloh2o.org/sara — addresses a crucial need in the Arabian Peninsula, providing a robust and reliable dataset to support hydrological modeling, water resource assessments, flood management, and climate research.

Copyright statement. TEXT

1 Introduction

The Kingdom of Saudi Arabia presents a striking hydrological paradox, experiencing periods of destructive flash floods and acute water scarcity, often simultaneously. Flash floods, the Kingdom's most frequent natural hazard, occur on average seven times a year across the country, incurring significant economic losses and social disruption (Al Saud, 2010). Particularly devastating were the flash floods in Jeddah in 2009 and 2011, claiming 113 and 10 lives, respectively, and resulting in widespread damage to property, averaging around 3 billion USD (Youssef et al., 2016). Moreover, climate change is projected to shift precipitation (P) patterns and increase atmospheric water vapor, potentially leading to more intense storms (Tabari and Willems, 2018; Almazroui et al., 2020; Fowler et al., 2021). Saudi Arabia also faces significant challenges in achieving water security for its growing population. The arid climate of the region, combined with an increasing water demand driven by rapid urbanization, industrialization, and agricultural expansion, puts immense pressure on limited water resources (Al-Ibrahim, 1991; Sultan et al., 2019). Effective management of these challenges requires accurate and timely P data, as well as to assess the impacts of climate change, develop adaptation and mitigation strategies, optimize water resources management, and improve flash flood early warning systems. The development of such datasets is also crucial for achieving the objectives of the National Water Strategy and Vision 2030 (Kingdom of Saudi Arabia, 2021), which aim to create a sustainable water sector while providing cost-effective supply and high-quality services to foster economic and social development

Over the past few decades, a wide range of gridded P products have been developed, each with unique design objectives, spatial and temporal resolutions, coverage, latency, algorithms, and data sources, ranging from satellite to analysis, reanalysis, gauges, and their combinations. Table 1 provides an overview of quasi- and fully-global products. In general, P products contain inherent errors and biases, making it important to assess their performance to determine their relative strengths, weaknesses, and suitability for different applications and utility for particular regions and geographies. While several global studies have assessed the performance of many of these products, typically using gauge P observations as reference (e.g., Beck et al., 2017; Sun et al., 2018; Nguyen et al., 2018), they have often excluded Saudi Arabia due to the scarcity of local P observations. To date, only a limited number of evaluations have specifically focused on the Arabian Peninsula (Kheimi and Gutub, 2015; Mahmoud et al., 2018; El Kenawy and McCabe, 2016; El Kenawy et al., 2019; Al-Falahi et al., 2020; Helmi and Abdelhamed, 2023; Alharbi et al., 2024; Jazem Ghanim et al., 2024). Two of these studies assessed individual satellite P products — IMERG (Mahmoud et al., 2018) and PDIR-Now (Alharbi et al., 2024) — leaving questions about the comparative performance of these products unresolved. Four other studies evaluated multiple satellite P products, including CMORPH, GSMaP, PER-SIANN, SM2RAIN-ASCAT, and TMPA 3B42 (Kheimi and Gutub, 2015; El Kenawy et al., 2019; Helmi and Abdelhamed, 2023; Jazem Ghanim et al., 2024). Notably, El Kenawy et al. (2019). Helmi and Abdelhamed (2023), and Jazem Ghanim et al. (2024) reported that the products generally performed poorly and highlighted the need for caution when using them. Al-Falahi et al. (2020) evaluated some of the aforementioned satellite P products as well as the reanalysis ERA5 but only focused on the highland region of Yemen. However, many satellite products evaluated in these studies have been superseded by newer, significantly improved versions. Additionally, several promising products have not been evaluated yet, including SM2RAIN-GPM (Massari et al., 2020), MSWEP V2.8 (Beck et al., 2019b), and JRA-3Q (Kosaka et al., 2024).

Several gridded P products, such as CHIRPS V2 (Funk et al., 2015), GPCP (Huffman et al., 2023), MSWEP V2.8 (Beck et al., 2019b), and SM2RAIN-GPM (Massari et al., 2020; Table 1), leverage multiple P-related data sources to obtain improved P estimates. These products employ statistical methods to minimize errors and biases inherent in individual sources, thereby enhancing P estimation performance across various regions, seasons, and temporal scales. Although these products generally outperform single-source P products (e.g., Beck et al., 2017; Prakash, 2019; Shen et al., 2020), machine learning (ML) approaches are increasingly recognized for their ability to efficiently fuse multiple data sources, while mitigating errors and biases. A wide variety of ML models, often trained with gauge observations, have been used for P estimation, including classical models such as multivariate linear regression (MLR), artificial neural networks (ANNs), support vector machines (SVMs), and random forests (RF) along with modern deep learning models such as convolutional neural networks (CNNs) and long short-term memory (LSTM) networks and hybrid models (see reviews by Hussein et al., 2022; Dotse et al., 2024; Papacharalampous et al., 2023; Xu et al., 2024). However, most ML studies on P estimation have limitations in that: (i) they generally focus on a small region or catchment, which limits the usefulness and generalizability of the findings; (ii) they often focus on a monthly (rather than daily or sub-daily) time scale, which may not meet the needs of all applications; (iii) they develop models for either near real-time or historical P purposes, but not both; (iv) they use gauge observations as predictors, which precludes near real-time model application, given that gauge observations are generally not available in near real-time; (v) they remain largely theoretical, often failing to offer a corresponding, accessible P dataset for users and follow-up studies. Additionally, and crucially, no study has yet investigated the potential of ML to specifically enhance P estimates in the Arabian Peninsula region.

Here, we introduce Saudi Rainfall (SaRa), a new gridded near real-time P product with an hourly $0.1°$ resolution designed to overcome the aforementioned limitations. The product covers the Arabian Peninsula (Figure 1) from 1979 to the present with a latency of less than 2 hours. It was derived using ML models trained on a vast database of hourly and daily gauge P observations from around the world. The ML models are tailored to various P product combinations to ensure optimal performance for each period and location. In the following section, we describe the data and methods underlying the product. Subsequently, we (i) evaluate the performance of the ML models, constructed using different gridded P product combinations as predictors, (ii) assess SaRa's performance relative to 19 global P products, (iii) examine the spatial patterns in performance, (iv) discuss the challenges of estimating P in arid regions, and (v) present trends in average and extreme P for the Arabian Peninsula based on SaRa.
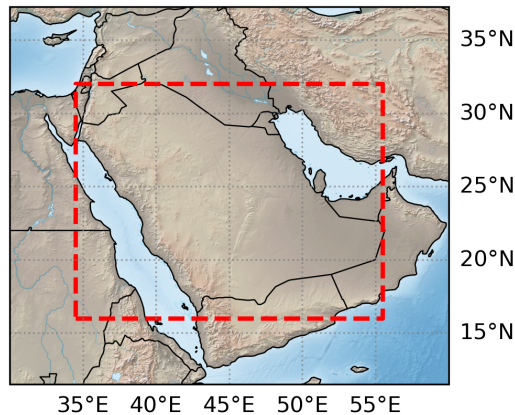
Figure 1. Study area location with SaRa dataset boundaries shown in red. Land cover data and shaded relief from Natural Earth (https://www.naturalearthdata.com/).

2 Data and Methods

2.1 Gridded precipitation and air temperature products

Global gridded P products (more details in Table 1) were used for two purposes: (i) as predictors to generate our new P product for the Arabian Peninsula (SaRa) and (ii) to evaluate the performance of SaRa relative to other P products. Gridded air temperature (T) data were also used as predictors to account for seasonal differences in error characteristics and relative performance among products. Throughout this paper, we refer to these global gridded predictors as "dynamic" due to their temporal variability — in contrast to predictors that are invariant in time, which are referred to as "static" (see below). We restricted our selection to P products with a daily or sub-daily temporal resolution. The P products included in our study originate from diverse sources, encompassing satellite observations, ground-based gauges, reanalyses, analyses, and combinations thereof. The ML models are trained to optimally merge the P products and mitigate errors and biases using gauge P data. The products selected as predictors for developing SaRa were primarily non-gauge-corrected P datasets to avoid biasing predictor importance, particularly in cases where the same stations used for training might have been employed to correct the respective P products. We included both microwave-based (IMERG-L V07 and GSMaP-MVK V8) and infrared-based (PERSIANN-CCS-CDR and PDIR-Now) satellite products as predictors. Among the predictors, the only gauge-corrected product is the infrared-based satellite-product PERSIANN-CCS-CDR, which has been corrected at the monthly scale using the Global Precipitation Climatology Project (GPCP) product (V2.3; monthly 2.5° resolution; Adler et al., 2018). PERSIANN-CCS-CDR was used prior to 2000, before the more accurate microwave-based satellite-products IMERG and GSMaP became available. For consistency, P and T estimates from ERA5 and GDAS were resampled from 0.25° to 0.1° using nearest neighbor to develop SaRa, while PERSIANN-CCS-CDR and PDIR-Now were resampled from 0.04° to 0.1° using averaging.

2.2 Static predictors

We used six static predictors to develop our new P product (Table 2). The term "static" indicates that these predictors are not time-dependent. Among these predictors, two are climate-related (Aridity Index, AI; and mean annual P, Pmean), one pertains to topography (Effective Terrain Height, ETH), and three are linked to geographic location (latitude, longitude, and absolute latitude; Lat, Lon, and AbsLat, respectively). AI represents the ratio of mean annual P to potential evapotranspiration (PET). ETH quantifies the orographic influence on P patterns by smoothing the topography (Daly et al., 2008). We excluded slope from our predictors as it was strongly correlated with ETH. Air temperature was omitted because it is already included as a dynamic predictor. Each static predictor was resampled by averaging to match the resolution of SaRa (0.1°).

2.3 Precipitation observations

We used P observations for two purposes: (i) as target to train the ML models underpinning our new P dataset, SaRa, and (ii) as reference to evaluate the performance of SaRa relative to other P products (Fig. 2). Although the SaRa product was specifically developed for the Arabian Peninsula, to account for the lack of hourly P data in the Arabian Peninsula, and to enhance the generalizability of the ML models, we trained the models using P observations from across the globe. This approach assumes that valuable insights from other regions can help optimize the merging of P products and reduce errors and biases in the Arabian Peninsula. Essentially, knowledge is transferred from data-rich (gauged) to data-poor (ungauged) regions, akin to regionalization techniques typically used in hydrology to tackle Predictions in Ungauged Basin (PUB) problems, which is considered a "grand challenge" in hydrology (Sivapalan et al., 2003; Hrachowitz et al., 2013).

For Europe and the conterminous US, our observational P data sources used for training and as reference were gridded P datasets based on gauge and radar data. Specifically, we used the EUropean RADar CLIMatology (EURADCLIM) dataset (hourly 2-km resolution; 2010–2022; Overeem et al., 2023) for Europe and the Stage-IV dataset (hourly 4-km resolution; 2002–present; Lin and Mitchell, 2005) for the conterminous US. To ensure the highest data quality, we extracted time series only at gauge locations from these datasets after resampling the data to the resolution of SaRa (0.1°) using averaging. We opted for these gauge-radar datasets over direct gauge observations in these regions because they: (i) provide grid-cell averages with probability distributions (e.g., peak magnitudes and P frequencies) matching those needed for SaRa (reducing the point-to-grid scale mismatch; Yates et al., 2006; Ensor and Robeson, 2008); (ii) are expressed in UTC, avoiding temporal shifts (Beck et al., 2019b; Yang et al., 2020); and (iii) have undergone extensive quality control (Lin and Mitchell, 2005; Overeem et al., 2023).

Outside Europe and the conterminous US, we used daily and hourly gauge P observations from various national, regional, and global data sources. The daily P data sources include: (i) the Global Historical Climatology Network-Daily (GHCN-D) dataset (ftp.ncdc.noaa.gov/pub/data/ghcn/daily/; Menne et al., 2012; 40,867 gauges), (ii) the Global Summary Of the Day (GSOD) dataset (https://data.noaa.gov; 9,904 gauges), (iii) the Latin American Climate Assessment & Dataset (LACA&D) dataset (http://lacad.ciifen.org/; 225 gauges), (iv) the Chile Climate Data Library (www.climatedatalibrary.cl; 712 gauges), and (v) national datasets for Brazil (10,963 gauges; www.snirh.gov.br/hidroweb/apresentacao), Mexico (3,908 gauges), Peru (255 gauges), Iran (3,100 gauges), and Saudi Arabia (459 gauges). The hourly P observations encompassed 2,312 gauges from the
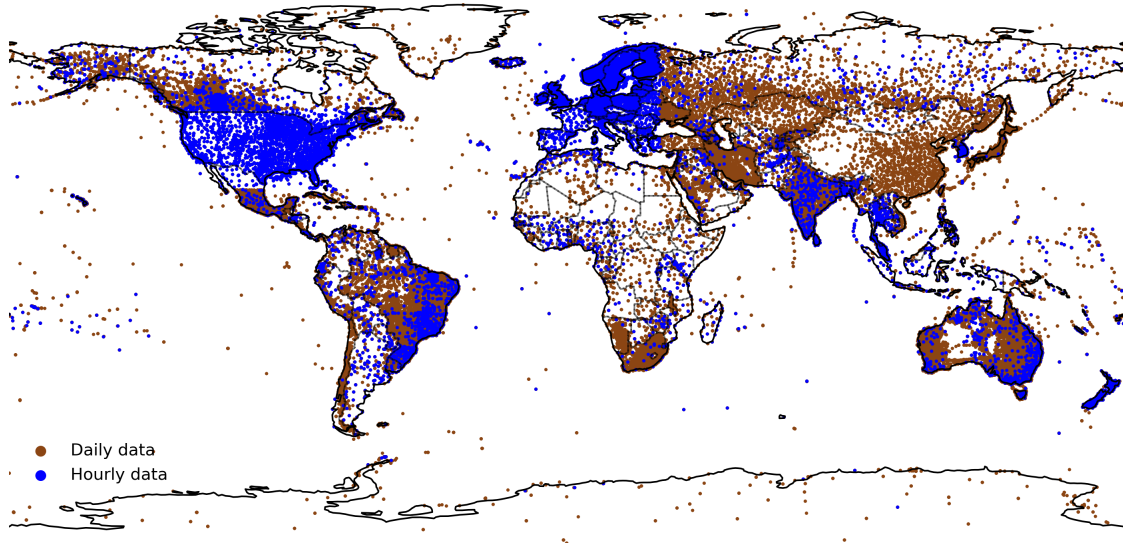
Figure 2. Global map of the 104,346 gauges remaining after duplication checks and quality control. Stations with hourly data are shown in blue, while stations with daily data are shown in brown.

Global Sub-Daily Rainfall (GSDR) dataset (Lewis et al., 2019) produced as part of the INTElligent use of climate models for adaptatioN to non-Stationary hydrological Extremes (INTENSE) project (Blenkinsop et al., 2018), 12,585 from the Integrated Surface Database (ISD) stations (Smith et al., 2011), and national gauges from Brazil (289 gauges).

The training and evaluation of the ML model stacks used to generate SaRa was carried out for the period 2010–2024. We used this period instead of the full 1979 to the present period to reduce the significant memory requirements associated with hourly data. Additionally, EURADCLIM data starts in 2010, most Saudi Arabian gauge records begin in 2014, and GDAS data starts in 2021.

2.4 Duplicates check and quality control

As we used P observations from a diverse range of data sources, there was an increased risk of some gauges being included in multiple sources. To avoid over-representation of these gauges in the training set and ensure the same data were not used for both training and evaluation, we removed these duplicates. To this end, we iterated over all gauges, and if another gauge was located within a 2-km radius, we gave preference to the source we deemed most reliable. We used the following order of most to least reliable source: EURADCLIM, Stage-IV, GHCN-D, GSDR, ISD, Bolivia, Brazil, Chile, Mexico, Iran, LACA&D, GSOD, and MEWA.

P observations are often subject to systematic, gross, and random errors (Kochendorfer et al., 2017; Tang et al., 2018), which can adversely affect the training and evaluation results. We identified and filtered out potentially erroneous gauges using the following five criteria: (i) non-zero minimum daily P; (ii) daily maximum less than 10 mm d^{-1} or exceeding 1,825 mm d^{-1} (the

highest daily rainfall ever recorded; www.weather.gov/owp/hdsc_world_record); (iii) mean annual P less than 5 mm yr^{-1} or exceeding 10,000 mm yr^{-1}; (iv) fewer than five P events (using a 1 mm d^{-1} threshold); and/or (v) fewer than 365 daily values (not necessarily consecutive) during 2010–2024 (the training and evaluation period). In total, 75,833 gauges were discarded out of 104,346 after these steps.

2.5 SaRa precipitation estimation algorithm

The SaRa product was derived using different ML model stacks trained from different combinations of dynamic P and T predictors (Section 1) along with several static predictors (Section 2). Each model stack is comprised of four separate ML submodels (Fig. 3). The first submodel is a daily XGBoost model (Chen and Guestrin, 2016) trained using daily P observations, leveraging the broad availability of P observations globally and in the Arabian Peninsula (Fig. 2). The second submodel, also based on XGBoost, disaggregates the daily estimates to 3-hourly and is trained using 3-hourly P observations, which are scarce in the Arabian Peninsula. As such, all of the 3-hourly dissagregation skill originates from other regions. As the resulting P estimates tend to underestimate the variance (i.e., generate excessive drizzle and underestimate peaks) due to the regression towards the mean phenomenon (see, e.g., He et al., 2016; Ting, 2024), a third submodel based on Random Forest (RF; Breiman, 2001) corrects the 3-hourly P probability distribution. The RF submodel is trained by separately sorting, for each gauge, (a) the 3-hourly estimates from the second submodel corrected using the daily estimates from the first submodel; and (b) the 3-hourly P observations. To ensure the number of wet days and low P intensities are also adequately corrected, the P estimates are square-root transformed before being fed to the third submodel, and the output of the third submodel is squared. The fourth submodel disaggregates the 3-hourly estimates to hourly and also represents an XGBoost model, trained using hourly P observations.

The dynamic predictors span different time periods and different regions (see Table 1), so we cannot use a single ML model stack for every time step and grid-cell. We therefore trained a total of 18 different ML model stacks with various combinations of dynamic predictors (Table 3). These model stacks are used based on the available dynamic predictors for a specific time and location (Figure 4), with preference given to the model stack with the lowest number (e.g., model_01 is preferred over model_02). The final SaRa P estimates were generated by iterating over all 0.1° grid-cells, loading all dynamic and static predictors, and then applying the preferred ML model stack. To avoid temporal discontinuities, such as around 1983 when PERSIANN-CCS-CDR was introduced or in 2000 when IMERG and GSMaP were introduced, the outputs from model_04 and model_05 were harmonized with the outputs of model_01, which we consider the reference due to its long record (from 2000 to 5 days prior to the present) and high accuracy (owing to the availability of ERA5, IMERG, and GSMaP). The harmonization process involved: (i) detrending the time series by dividing by the moving annual average, (ii) cumulative distribution function (CDF) matching, and (iii) multiplying the result by the moving annual average (Figure 4). The detrending serves to avoid amplification of trends in extreme P (see, e.g., Cannon et al., 2015).

We implemented the RF models using the `scikit-learn` package and XGBoost models using the `XGBoost` package in Python. The hyperparameters we used are summarized in Appendix A.
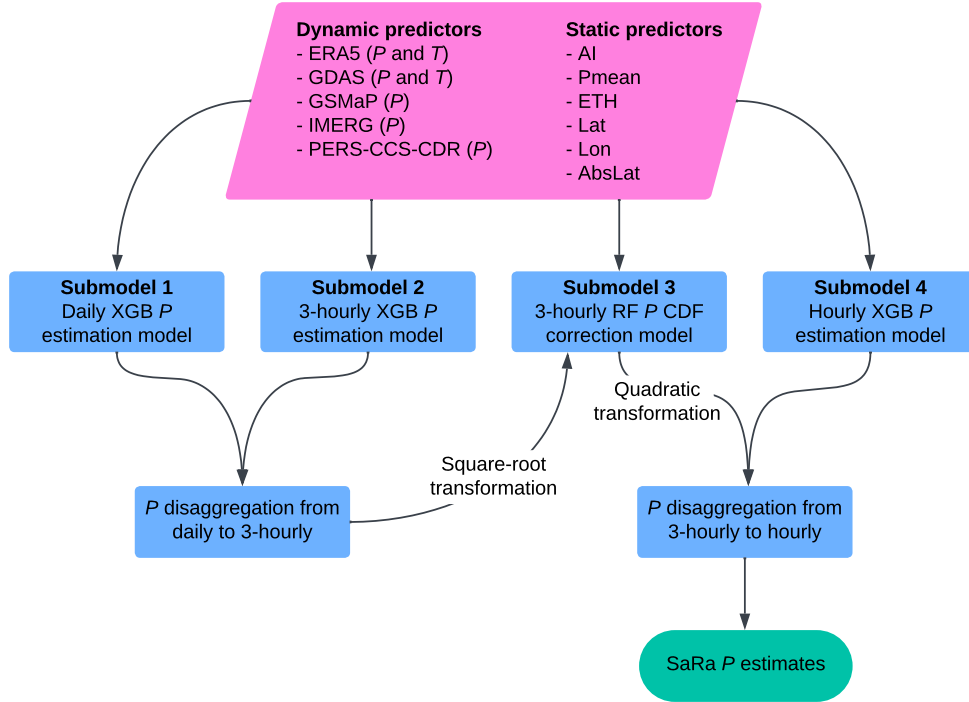
Figure 3. Flowchart of ML model stacks used to produce the new SaRa P product presented in this study.

2.6 Training and evaluation

Both the training and evaluation were carried out for the period 2010–2024, aligning with the temporal coverage of EURAD-CLIM and the Saudi gauge data. From the 28,513 gauges that passed the duplicates check and quality control (Section 2.4), we randomly allocated 50 % for training (14,256 gauges) and the remaining 50 % for evaluation. We trained submodel one using all available daily gauge data, while submodels two, three, and four were trained using gauges with hourly data (aggregated to 3-hourly for submodels two and three). The specific number of gauges used for training each model stack depends on the temporal span and spatial coverage of the dynamic predictors. Since GDAS covers a relatively short period (2021 to the present), any ML models incorporating GDAS were trained using a significantly smaller amount of observations.

The time stamps of the hourly P observations may reflect the local time zone instead of Coordinated Universal Time (UTC), while the daily P observations may represent accumulations ending at various times, not necessarily at midnight UTC (Yang et al., 2020). Such discrepancies can lead to temporal mismatches between the dynamic predictors on one side and the gauge P observations on the other, thereby hindering satisfactory model training. To address this issue, we determined time shifts in the hourly and daily gauge P data using the hourly satellite-based IMERG-L V07 and GSMaP-MVK V8 products, similar to Beck et al. (2019b). To determine hourly gauge data shifts, we shifted the gauge record by 1-hour increments from -36 to $+36$ and calculated the average Spearman correlations between shifted gauge data and IMERG-L V07/GSMaP-MVK V8

Figure 4. (a) Different ML model stacks were used for different periods and locations to account for differences in the spatio-temporal availability of the dynamic predictors. The primary ML model stack, the corresponding dynamic predictors, and the mean daily independent validation KGE (from Table 4) are also provided. Note that other ML model stacks may be used for a particular period when any dynamic predictor is not available. (b) Conceptual illustration of how P estimates from different ML model stacks are combined, and how outputs from ML model stacks with long records are harmonized with the reference (model_01).

data. To determine daily gauge data shifts, we shifted the IMERG-L V07 and GSMaP-MVK v8 data separately by 1-hour increments from -36 to $+36$, computed daily P accumulations from the shifted IMERG-L V07 and GSMaP-MVK V8 data, and calculated the average Spearman correlations between daily gauge record and shifted IMERG-L V07/GSMaP-MVK V8 data. The shifts that yielded the highest correlations were then used to recalculate daily values of the dynamic predictors for 200 training ML submodel one, as well as to shift the hourly gauge records for training submodels two and four.

The 119 randomly selected evaluation gauges are completely independent and were not used to train the ML models, enabling a thorough performance assessment of SaRa compared to other P products (Table 4). We used several performance metrics for a comprehensive evaluation: (i) the Kling-Gupta Efficiency (KGE; Gupta et al., 2009; Kling et al., 2012), which is an aggregate metric combining Pearson correlation (r_dly), overall bias (β), and variance bias (γ), (ii) monthly Pearson correla-205 tion (r_mon), (iii) peak bias at the 99.5th percentile (B_peak; %), (iv) wet day bias ($B_\mathrm{wet\,days}$, days; calculated using a 0.5 mm d^{-1} threshold), and (v) Critical Success Index (CSI), measuring the ratio of hits to the sum of hits, false alarms, and misses for

P events exceeding 10 mm d^{-1}. These metrics, selected to encompass all important aspects of P time series, were computed for each evaluation gauge based on daily data, except for r_{mon}. We did not conduct an hourly evaluation due to the lack of hourly observations in the Arabian Peninsula. For a detailed explanation of the performance metrics, including the equations,
210   see Appendix B.

3   Results and Discussion

3.1   Performance of ML models

To generate the SaRa product, we trained 18 different ML model stacks with different P product combinations (Table 3). The performance of these models in terms of median KGE, calculated using daily P data from independent evaluation gauges,
215   ranges from 0.03 for model_05, which relies solely on one dynamic P predictor (ERA5), to 0.43 for model_06, which incorporates four dynamic P predictors (ERA5, GDAS, IMERG-L V07, and GSMaP-MVK V8; Table 4). These results align with our expectation that models incorporating a larger number of dynamic predictors are able to extract complementary strengths from them, resulting in better performance. model_01, based on three dynamic P predictors (ERA5, IMERG-L V07, and GSMaP-MVK V8), is arguably the most important ML model stack as it covers the largest portion of the record, from 2000 to
220   5 days prior to the present, also performs well, achieving a median KGE of 0.36, a median peak bias (B_{peak}) of -11.17 %, a wet day bias ($B_{\mathrm{wet\,days}}$) of $+1.42$ days, and a median weather event detection score (CSI$_{10\,\mathrm{mm}}$) of 0.21.

The B_{peak} value of -11.17 % obtained by model_01 suggests a slight underestimation of high rainfall amounts (Table 4). Although ML models are known to underestimate extremes due to the regression toward the mean phenomenon (see, e.g., He et al., 2016), the most likely reason is that the gauge P data used for evaluation represent point measurements, which typically
225   exhibit higher peaks than grid-cell averages (Ensor and Robeson, 2008). Thus, this apparent underestimation may primarily reflect a scale discrepancy. Similarly, the $B_{\mathrm{wet\,days}}$ value of $+1.42$ days (Table 4), indicating a minor overestimation of rainfall frequency, might also be attributable to this scale difference. Rainfall frequencies are generally higher for grid-cell averages than for point measurements (Osborn and Hulme, 1997). Note that a significant portion of the training data comprises gridded gauge-radar P data (see Section 2.3), which have a 0.1° grid-cell scale consistent with SaRa.
230   Among all the trained ML model stacks, model_05 performed the worst, with a median KGE of 0.05. model_05 uses only one dynamic predictor (ERA5) and is designed primarily for the period before 1983, when only ERA5 data are available (Figure 4). It also performs worse than ERA5 alone (median KGE of 0.21), which is mainly attributable to poor bias (β) and peak bias (B_{peak}) values. Fortunately, these issues are largely resolved during the harmonization step, where the outputs of model_04 and model_05 are harmonized with those of model_01, which is considered the reference (Figure 4).
235   A key potential limitation of ML-based P estimation is poor generalizability; models often fail in regions lacking training data (Xu et al., 2024). To assess whether this applies to our models, we analyzed KGE values of the evaluation stations as a function of distance to the nearest training station (Figure 5). The results show no clear decline in KGE with increasing distance, indicating satisfactory spatial generalizability. Another potential limitation is the "black-box" nature of ML models, which limits interpretability. To improve transparency, we computed predictor importances for all four submodels of model_01
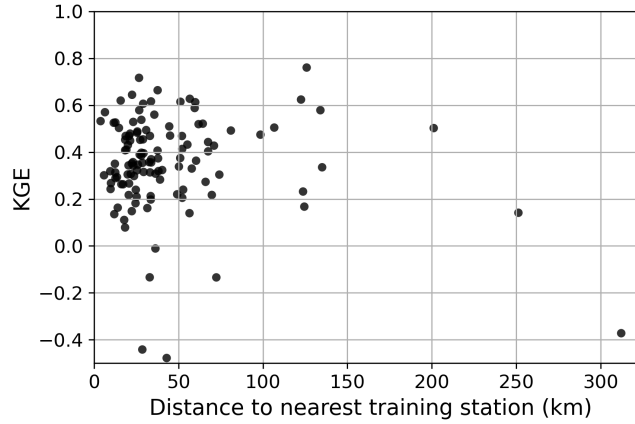
Figure 5. KGE versus distance to nearest training station for the evaluation stations ($n = 119$). The same KGE values are also presented in Fig. 6a.

240   (Table 6). IMERG-L V07 consistently ranked higher than GSMaP-MVK V8 in importance, indicating a preference for IMERG, in agreement with its superior validation performance (Table 4). ERA5 was the most important predictor for the daily submodel (Submodel 1), whereas IMERG dominated in the 3-hourly and hourly submodels (Submodels 2 and 4, respectively). This likely reflects the ability of observational satellite-based datasets like IMERG to capture event timing more accurately. This also demonstrates the ability of the models to exploit the complementary strengths of the different P predictors. Static predictors

245   were overall much less important than dynamic ones. Among the static predictors, Lon, Lat, and AbsLat had the highest importances, accounting for regional variability in P predictor performance and error characteristics.

3.2   Performance comparison with other gridded P products

The primary model underlying our newly developed SaRa product, model_01, exhibited superior performance across nearly all 12 performance metrics relative to all 19 other P products (Table 4). Notably, model_01 attained a median KGE of 0.36,

250   significantly outperforming widely used P products such as ERA5, JRA-3Q, CMORPH-RT, CHIRPS V2, IMERG-L V07, GSMaP MVK V8, and MSWEP V2.8, which obtained median KGE values of 0.21, 0.12, 0.21, 0.17, -0.39, -0.13, and 0.2, respectively. Additionally, model_01 performed well in terms of high P intensities, exhibiting a low peak bias (B_{peak}) of -11.17%, while the aforementioned other P products showed higher biases of -30.18%, -18.02%, -44.55%, -27.82%, $+72.91\%$, $+44.7\%$, and -28.65%, respectively. In terms of detecting P events ($\text{CSI}_{10\,\text{mm}}$), model_01 scored a median value

255   of 0.22, surpassing the other products with values ranging from 0.09 to 0.19.

    Although SaRa was derived using an algorithm trained on gauge P observations (from stations excluded in the evaluation), it was not directly corrected using gauge observations. Despite this, SaRa outperformed products entirely based on gauge observations (CPC Unified) or corrected using gauge observations (CHIRPS V2, IMERG-F V07, and MSWEP V2.8). This mainly reflects the limited availability of gauge observations from Saudi Arabia in global databases like GHCN-D (Menne

et al., 2012; Kidd et al., 2017). Additionally, the lower performance of CHIRPS V2 and IMERG-F V07 may stem from their 5-day and monthly gauge corrections, respectively, which are less effective at improving performance on a daily time scale.

Among the purely (re)analysis-based products (ERA5, GDAS, and JRA-3Q), GDAS performed best with a median KGE of 0.24, outperforming ERA5 (median KGE of 0.21) and JRA-3Q (median KGE of 0.12; Table 4). Among the microwave satellite-based products (CMORPH-RT and -RAW, GSMaP-MVK V8, IMERG-L and -E V07, and SM2RAIN-GPM), CMORPH-RT emerged as best (median KGE of 0.21), followed by SM2RAIN-GPM (median KGE of 0.18), and GSMaP-MVK V8 (median KGE of -0.13). Among the purely infrared satellite-based products (PERSIANN-CCS and PDIR-Now), PDIR-Now performed best (median KGE of 0.11). PDIR-Now also outperformed some microwave-based products (GSMaP-MVK V8 and IMERG-L and -E V07). Among the SM2RAIN products (SM2RAIN-GPM, SM2RAIN-ASCAT, and SM2RAIN-CCI), SM2RAIN-GPM obtained the best overall performance (median KGE of 0.18). SM2RAIN-ASCAT and SM2RAIN-CCI performed poorly, exhibiting $CSI_{10\,mm}$ values of 0.00, underscoring the limited capability of algorithms that infer P based on soil moisture signals to detect P events > 10 mm d^{-1} over the Arabian Peninsula. This is likely due to the extremely arid conditions in the Arabian Peninsula, where the soil dries out rapidly following P events, reducing the effectiveness of soil moisture-based detection methods. These results are in agreement with Jazem Ghanim et al. (2024), who also found SM2RAIN-ASCAT to perform poorly in the Arabian Peninsula.

Previous studies that evaluated P datasets for the Arabian Peninsula include Kheimi and Gutub (2015), Mahmoud et al. (2018), El Kenawy and McCabe (2016), El Kenawy et al. (2019), Al-Falahi et al. (2020), Helmi and Abdelhamed (2023), Alharbi et al. (2024), and Jazem Ghanim et al. (2024). Comparing our results to these studies is challenging because they assessed fewer P products and used outdated versions. However, Alharbi et al. (2024) reported a mean r_{dly} of 0.33 for PDIR-Now, which is comparable to our median value of 0.32 (Table 4). Similarly, Helmi and Abdelhamed (2023) reported KGE, r_{dly}, r_{mon}, and $CSI_{10\,mm}$ values for PERSIANN-CCS-CDR, CHIRPS, and IMERG-F that are also consistent with our results.

3.3 Spatial distribution of performance metrics

To illustrate SaRa's performance, Fig. 6 shows the spatial distribution of its performance relative to other widely used P datasets. Figures 6a and Figure 6b display the spatial distribution of SaRa's KGE and $CSI_{10,mm}$, respectively, as obtained by SaRa (model_01) during its evaluation over independent rain gauges across Saudi Arabia. The differences in performance between SaRa and three widely used global products (CHIRPS V2, ERA5, and IMERG-L V07) are also shown for both metrics (Figs. 6c–h), highlighting SaRa's overall superior performance. At first glance, the spatial distribution of both metrics appears random, with clusters of good performance adjacent clusters of poor performance, lacking a clear spatial organization. This randomness may be partly due to rain gauge measurement errors (Ciach, 2003; Daly et al., 2007; Sevruk et al., 2009), compounded by scale discrepancies between point-scale measurements and grid-scale averages (Yates et al., 2006).

To examine whether performance patterns are related to specific climatic or topographic factors, we calculated Spearman rank correlation coefficients between climatic and topographic attributes of the evaluation rain gauges (Table 2) and the performance scores (Section 2.6) for SaRa's model_01, SaRa's model_06, MSWEP V2.8, IMERG-L V07, GSMaP MVK V8, and ERA5 (Table 5). Overall, the correlations were slightly weaker for model_01 and model_06 compared to the P products,

suggesting more stable performance, which is expected, given that (i) the ML models leverage the complementary strengths of the P predictors and (ii) the climatic and topographic variables were included as predictors in the models.

The metric r_{dly}, which evaluates the ability of models or products to estimate daily P variability, is primarily sensitive to random errors and less influenced by systematic biases, which are relatively easy to correct. In all models and products, r_{dly} shows a positive correlation with aridity index (AI), indicating reduced performance in arid regions. This conforms with previous large-scale P product evaluations (e.g., Beck et al., 2017; Sun et al., 2018; Abbas et al., 2025) and reflects the brief, intense, and localized nature of rainfall in such regions. Additionally, r_{dly} exhibited negative correlations with effective terrain height (ETH) for all models and products, indicating lower performance in the mountainous southwest where orographic P predominates. This is also consistent with other P product evaluations (e.g., Ebert et al., 2007; Derin et al., 2016; Beck et al., 2019a) and reflects the greater heterogeneity of P in regions of complex terrain. Additionally, satellite retrieval of P in mountainous regions is particularly challenging due to the shallow nature of orographic P (Yamamoto et al., 2017; Adhikari and Behrangi, 2022).

Performance metrics related to systematic biases in magnitude (β and B_{peak}) generally showed negative correlations with mean precipitation (Pmean) and positive correlations with the aridity index (AI). These correlations were particularly strong for IMERG-L V07, suggesting that this product could benefit from bias correction using climate indices.

3.4 Challenges of precipitation estimation in arid regions

Although SaRa outperformed all other P products, its performance metrics might seem underwhelming. For instance, the daily Pearson correlation (r_{dly}) of 0.50 achieved by model_01 (SaRa's primary model; Table 4) indicates that only 25% (100×0.50^2) of the daily variability in P observations is captured, while other products perform worse. Similarly, a $\mathrm{CSI}_{10\,\mathrm{mm}}$ of 0.21 suggests a moderate ability to detect P events exceeding 10 mm d^{-1}. These results align with prior large-scale evaluations reporting lower accuracy of P products in arid regions (e.g., Beck et al., 2017; Sun et al., 2018; Abbas et al., 2025), reflecting the inherent challenge of precipitation estimation in these environments.

The challenges in arid regions stem from several key factors:

1. P events in arid regions are typically short-duration, highly intense, and spatially localized, making them difficult to detect with satellites, simulate with models, or measure with gauges. This contrasts with temperate and cold climates, where P events are generally longer-lasting, less intense, and spatially broader (El Kenawy et al., 2019; Ebert et al., 2007).

2. Reanalyses like ERA5 and JRA-3Q and analyses like GDAS are based on numerical weather prediction (NWP) models, which struggle to simulate the complex convective processes prevalent in arid regions, including deep convection initiation, rapid cell dissipation, and the effects of dry boundary layers (Yano et al., 2018; Peters et al., 2019; Lin et al., 2022).
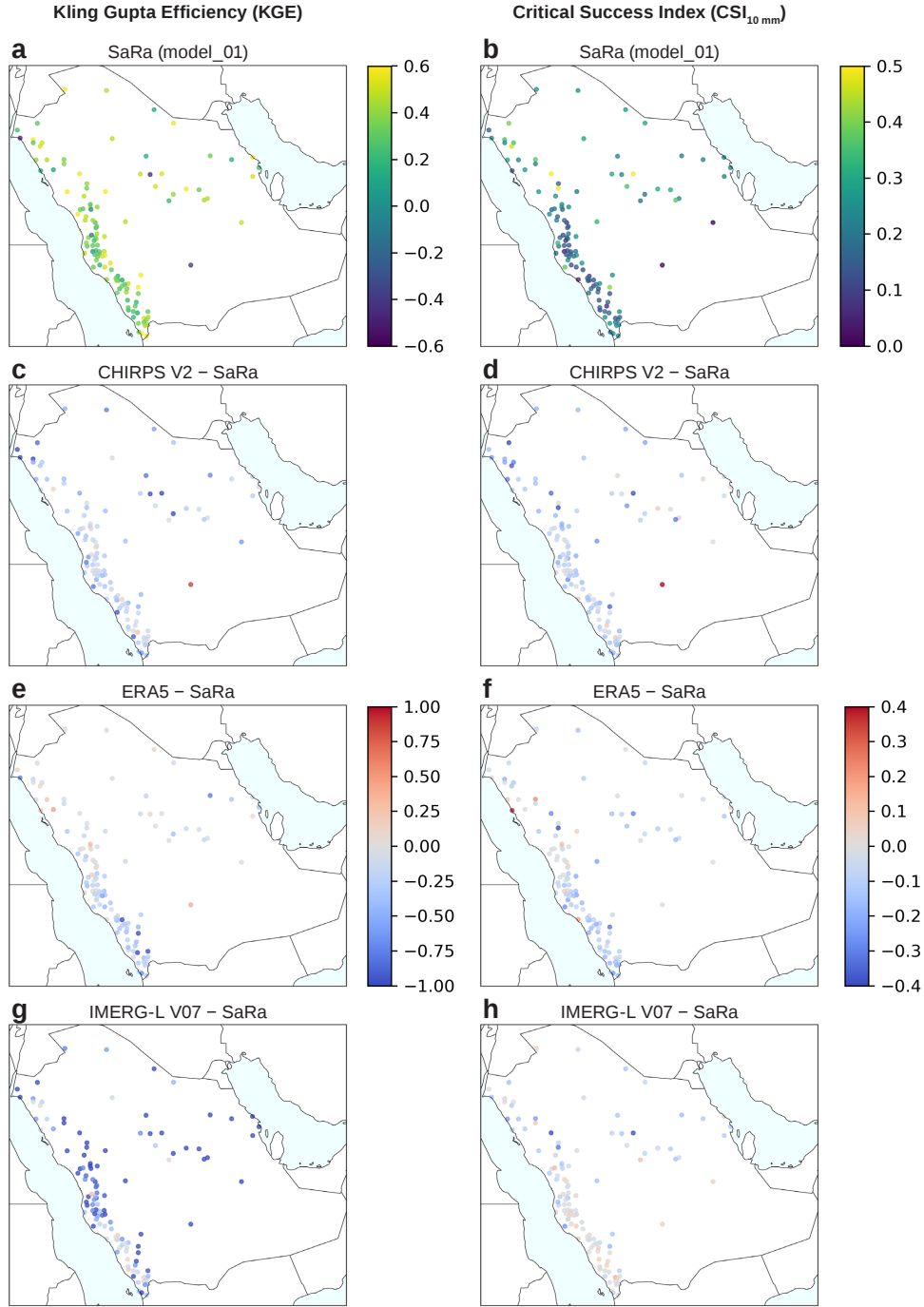
Figure 6. Performance of the primary SaRa ML model (model_01) in terms of (a) Kling-Gupta Efficiency (KGE) and (b) detection of P events ($CSI_{10\,mm}$). (c,e,g) difference in KGE between model_01 and CHIRPS V2, ERA5, and IMERG-L V07, respectively. (d,f,h) difference in $CSI_{10\,mm}$ between model_01 and CHIRPS V2, ERA5, and IMERG-L V07, respectively. Each data point represents an independent evaluation gauge ($n = 119$).

3. Virga P, which evaporates before reaching the ground, adds another layer of complexity. It is estimated to account for nearly half of all events in some arid regions, leading to significant false detections by satellite radiometers (Wang et al., 2018).

4. Errors in gauge measurements — due to wind deflection, evaporation within the gauge, splashing, and wetting losses — can also play a significant role (Ciach, 2003; Daly et al., 2007; Sevruk et al., 2009). In arid regions, higher evaporation rates exacerbate wetting losses, and the sparse, short-lived nature of rainfall events amplifies sampling errors (Villarini et al., 2008).

5. Discrepancies between point measurements from gauges and grid-cell averages derived from satellite or model products also contribute to lower performance scores (Yates et al., 2006; Ensor and Robeson, 2008). In arid regions, this mismatch is likely particularly significant, due to the highly localized nature of P.

6. Time shifts between daily P totals from gauges and satellite or (re)analysis products (Yang et al., 2020; Beck et al., 2019b) further reduce performance scores, especially in arid regions due to the short duration of rainfall events. The boundary between daily totals from satellite or (re)analysis products is midnight UTC, whereas it varies for daily gauge totals depending on regional reporting practices. In Saudi Arabia, the average boundary time was determined to be 05:00 AM UTC (08:00 AM local time; see Section 2.6). Consequently, for a brief event of one hour, there is a $100 \times 5/24 = 21$ % chance that it will be assigned to the 'wrong' day.

7. In arid regions, where rainfall is infrequent and therefore measurements are often considered unnecessary, the number of stations is usually limited (Menne et al., 2012; Kidd et al., 2017). As a result, P products may not be evaluated in these areas during development, potentially leading to lower performance.

However, it should be kept in mind that a hypothetical baseline P product predicting only the mean would achieve a KGE of -0.41 (Knoben et al., 2019), making SaRa's model_01 median KGE of 0.36 quite reasonable, situated between this baseline and an (unattainable) perfect score of 1. Furthermore, performance improves markedly when data is averaged over longer periods, as evidenced by the median monthly Pearson correlation (r_{mon}) of 0.71 for model_01 (Table 4), indicating that more than twice as much variability is captured at the monthly scale compared to the daily scale. Similarly, performance is enhanced at larger spatial scales, for example when computing regional averages or when driving a hydrological model for a catchment. This improved performance reflects the reduced impact of errors due to spatial aggregations within catchments and regions (O and Foelsche, 2019).

3.5 Precipitation climatology and trends in Saudi Arabia

According to the newly developed SaRa P product, the mean annual P for Saudi Arabia during the period 1991–2020 is 54 mm yr^{-1} (Fig. 7a). However, although SaRa's median β score is 1.03 (Table 4), indicating negligible bias according to this metric, the gauge-based mean (rather than median) P is nonetheless 18 % higher than the SaRa-based mean P across all evaluation gauges, likely reflecting the tendency of ML models to attenuate extremes (see, e.g., He et al., 2016; Ting,

2024). Consequently, an adjusted estimate of $54 \times 1.18 = 64$ mm yr^{-1} may represent the best estimate for mean annual P in Saudi Arabia. This value is significantly lower than the estimate of 102 mm yr^{-1} for the period 1991–2020 from the Climatic Research Unit (CRU) gridded Time Series (TS) dataset (Harris et al., 2020), as published on the World Bank website (https://climateknowledgeportal.worldbank.org/country/saudi-arabia/climate-data-historical). However, the CRU climatology is based on interpolation of a small number of gauges (approximately ten) in Saudi Arabia (New et al., 1999), whereas SaRa was trained using 113 stations in the country. The mean annual P estimate of 84 mm yr^{-1} from Almazroui (2011), based on TMPA 3B42 (Huffman et al., 2007) bias-corrected using 29 stations, comes closer to our estimate, although it covers a different period (1998–2009).

In addition to annual totals, SaRa provides insights into P frequency and peak P magnitudes. The average annual maximum daily P in Saudi Arabia is 19 mm d^{-1} (Fig. 7c). While this value is modest compared to the global best-estimate mean of 56 mm d^{-1} (from the global gauge-based Precipitation Probability DISTribution — PPDIST — dataset V1.0; Beck et al., 2020), such P extremes can nonetheless cause severe flooding due to the region's low soil infiltration capacity, sparse vegetation, and insufficient flood management infrastructure (Othman et al., 2023). The average annual maximum hourly P is 6.9 mm h^{-1}. On average, Saudi Arabia experiences 10 rainy days per year (defined as days with $P \geq 0.5$ mm d^{-1}; Fig. 7e) and 51 rainy hours per year (defined as hours with $P \geq 0.1$ mm h^{-1}). For context, the global average number of rainy days per year is 30, based on PPDIST, using the same threshold of 0.5 mm d^{-1}. Across all metrics — mean annual P, P frequency, and P extremes — the highest values occur along the western slopes of the Asir Mountains, where orographic effects enhance P (Hasanean and Almazroui, 2015).

Trend analysis from 1979 to 2023 based on SaRa reveal declines in mean annual P, daily P frequency, and annual maximum daily P at rates of -0.50, -0.11, and -0.58 % yr^{-1}, respectively (see Figs. 7b, 7d, and 7f, respectively). Over the 45-year period, these rates correspond to cumulative reductions of -22.5 %, -5.0 %, and -26.1 %, respectively. Our results align with Almazroui (2020), who reported a mean annual P trend of -0.65 % yr^{-1} during 1978–2019 based on 25 stations in Saudi Arabia. They are also consistent with Munir et al. (2025), who analyzed Standardized Precipitation Index (SPI) time series from 28 stations in Saudi Arabia for 1985–2023, finding negative trends at 16 stations and positive trends at 10. Furthermore, the strong declines observed in southeastern Saudi Arabia (Fig. 7b) align with Patlakas et al. (2021), who analyzed trends for 1986–2015 using a regional atmospheric model. These trends result from multiple factors, including internal climate variability, external natural influences, and human-induced climate change.

However, it is important to note that these trend estimates are mostly statistically insignificant (p-value > 0.05) and subject to substantial uncertainty due to large interannual variability, as well as considerable errors in gauge, reanalysis, and satellite P estimates (see Section 3.4). In addition to random errors, satellite datasets are affected by transitions in data sources and radar sensors used for calibration (e.g., TRMM to GPM circa 2015; see Huffman, 2019), while reanalyses are affected by updates in data assimilation, such as the progressive inclusion of new satellite datasets (e.g., the TOVS to ATOVS transition in 2000), as well as the concatenation of different production streams (Hersbach et al., 2020). These discontinuities propagate through and are reflected in SaRa, contributing to the observed uncertainties and hindering the detection of significant trends. Despite historical declines, future projections from climate models in the sixth phase of the Coupled Model Intercomparison
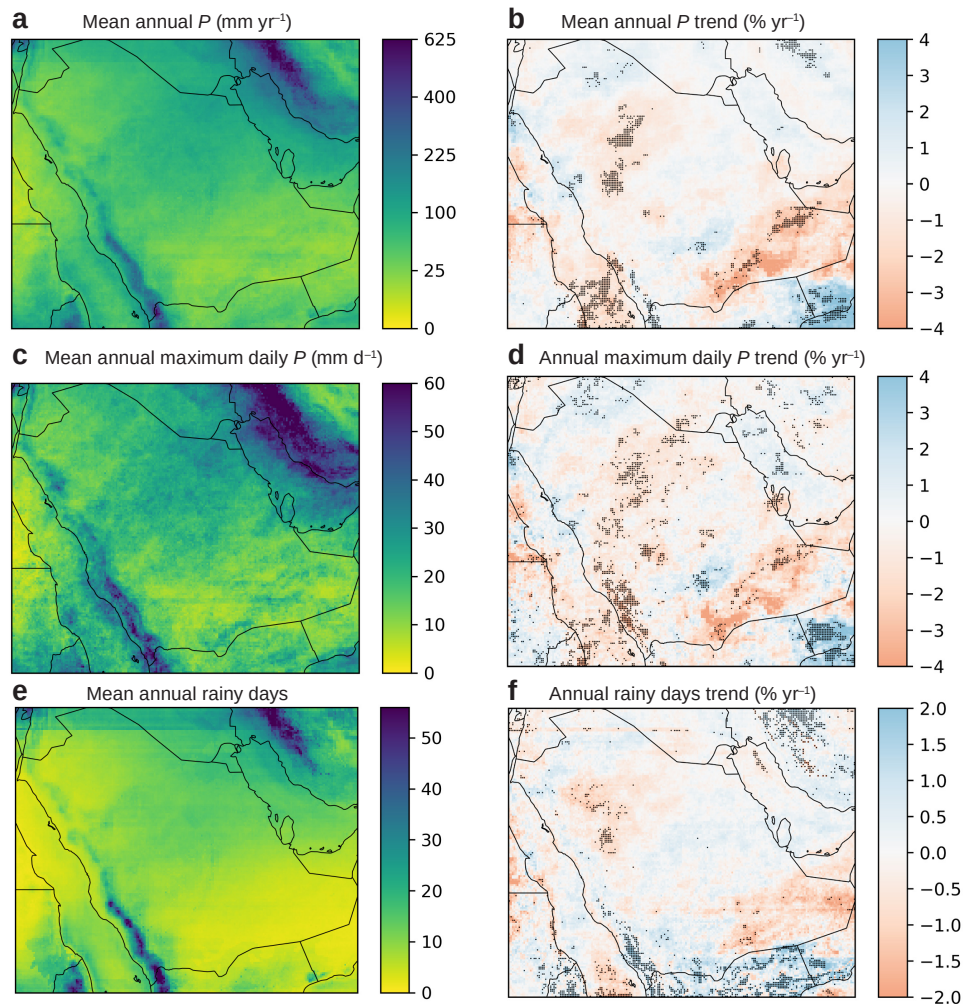
Figure 7. Mean and trend values (1979–2023) based on the new SaRa P product for (a–b) mean annual P, (c–d) annual maximum daily P, and (e–f) annual number of wet days (using a threshold of 0.5 mm d^{-1}). Grid cells with statistically significant trends (p-value < 0.05) are marked with black dots. Note that panel (a) uses a non-linear color scale.

Project (CMIP6) indicate that increases in all three metrics (mean annual P, annual maximum daily P, and annual rainy days) are likely across most regions of Saudi Arabia (Iturbide et al., 2021; Intergovernmental Panel on Climate Change (IPCC), 2023), highlighting the need and value of data-driven P approaches in resolving potential discrepancies in P distributions and spatio-temporal patterns.

4 Conclusions

The SaRa dataset, a high-resolution, gridded and near real-time P product was developed to satisfy the critical need for more accurate and robust P data in the Arabian Peninsula. SaRa offers hourly data at a 0.1° resolution spanning from 1979 to the present, with a latency of less than two hours. The algorithm underpinning the product involves 18 ML model stacks tailored to various P product combinations to ensure optimal performance for each period and location. These models were trained using daily and hourly gauge P observations from across the globe to enhance their accuracy. Our primary findings are summarized as follows:

1. Among the 18 model stacks, model_06, which incorporates four dynamic predictors (ERA5, GDAS, IMERG-L V07, GSMaP-MVK V8), achieved the highest median KGE (0.43). model_01, the primary model spanning the longest period (2000 to five days before the present) and incorporating three dynamic predictors (ERA5, IMERG-L V07, GSMaP-MVK V8), also demonstrated strong performance (median KGE of 0.36). Most models showed minimal bias in peak rainfall and wet day frequency, reinforcing the idea that product performance can be enhanced by leveraging the complementary strengths of diverse P datasets. However, model_05, relying solely on ERA5, performed relatively poorly (median KGE of 0.03) due to significant bias issues, although these were mitigated through harmonization with model_01.

2. We carried out the most comprehensive daily evaluation of gridded P products in the Arabian Peninsula to date. SaRa outperformed all 19 other gridded P products across nearly all 12 performance metrics, achieving a median KGE of 0.36. Notably, it demonstrated superior event detection and lower peak bias compared to state-of-the-art products such as ERA5, JRA-3Q, CMORPH, and IMERG-F V07. Among the evaluated products, microwave-based satellite products generally performed better than infrared-based satellite products and (re)analyses.

3. The spatial performance analysis of SaRa's model_01 across Saudi Arabia, based on KGE and $CSI_{10,mm}$ metrics, reveals a seemingly random distribution of performance, with clusters of high and low performance influenced by rain gauge errors and scale discrepancies between point and grid measurements. Correlations with climatic and topographic variables suggest relatively stable performance, as these variables were incorporated as predictors in the model. Performance in estimating daily P variability (r_{dly}) decreases in arid regions and mountainous areas, reflecting challenges with localized, intense rainfall and shallow orographic P.

4. Despite outperforming all other products, SaRa's performance may nonetheless seem somewhat underwhelming, highlighting the inherent difficulty in estimating P in arid regions, where events are typically localized, brief, and intense. For example, SaRa's primary model, model_01, captured only 25 % of daily P variability, and the other P products less, possibly due to challenges including virga, short-duration and highly variable rainfall. However, it is worth noting that achieving perfect scores is impossible due to inherent gauge errors, scale discrepancies, and time shifts in daily accumulations.

5. Mean annual P across Saudi Arabia was estimated as 64 mm year^{-1} over the period from 1991–2020, which is significantly lower than prior estimates based solely on rain gauges. Saudi Arabia averages 10 rainy days and 51 rainy hours

annually, with higher and more frequent P in the southwestern Asir mountains due to orographic effects. From 1979 to 2023, P trends show a decline in annual totals, frequency, and extremes (up to -26.1 %), driven by climate variability and anthropogenic factors. In contrast to the historical trends, climate projections suggest potential future increases.

Our study addresses the long-standing need for more accurate P estimates and provides a comprehensive evaluation of gridded P products in Saudi Arabia, one of the most arid, water-stressed, and data-sparse regions on Earth. The SaRa dataset, available at www.gloh2o.org/sara, equips researchers, professionals, and policymakers with the tools needed to tackle pressing environmental and socio-economic challenges in Saudi Arabia, and serves not only as a potential framework for filling this data gap in other arid and dryland regions, but as a framework that could be applied globally to develop a consistent long-term dataset. The product delivers a high-resolution, near real-time resource designed to support a diverse range of applications, including water resource management, hydrological modeling, agricultural planning, disaster risk reduction, and climate studies.

Code availability. The code used to generate the figures is available from the corresponding author upon request.

Data availability. CPC Unified is available from the NOAA NOAA Physical Sciences Laboratory (PSL) website (https://psl.noaa.gov/data/gridded/data.cpc.globalprecip.html). IMERG V07 is accessible from the NASA Global Precipitation Measurement (GPM) website (https://gpm.nasa.gov/data). JRA-3Q is available through the National Center for Atmospheric Research (NCAR) Research Data Archive (RDA; https://rda.ucar.edu/datasets/ds640000/dataaccess). SM2RAIN-ASCAT, SM2RAIN-CCI, and GPM+SM2RAIN are hosted on Zenodo (https://zenodo.org/records/10376109, https://zenodo.org/records/1305021, and https://zenodo.org/records/3854817, respectively). ERA5 data can be obtained from the Copernicus Climate Data Store (CDS; https://cds.climate.copernicus.eu/datasets/reanalysis-era5-single-levels?tab=overview). CHIRP and CHIRPS V2 are available via the University of California, Santa Barbara, Climate Hazards Center (CHC) website (https://www.chc.ucsb.edu/data/chirps/). MSWEP V2.8 is accessible from the GloH2O website (https://www.gloh2o.org/mswep/). PERSIANN-CCS-CDR and PDIR-Now are available from the Center for Hydrometeorology and Remote Sensing (CHRS) website (https://chrsdata.eng.uci.edu/). CHELSA is accessible at https://chelsa-climate.org.

Appendix A: ML model hyperparameters

The hyperparameters used for the RF and XGBoost models are described in Tables A2 and A1, respectively. These hyperparameters were selected to balance model complexity and training time, while also minimizing the risk of overfitting.

Appendix B: Performance metrics calculation

The Kling-Gupta Efficiency (KGE) is given by:

$$\text{KGE} = 1 - \sqrt{(r-1)^2 + (\beta-1)^2 + (\gamma-1)^2}, \tag{B1}$$

where r is the Pearson correlation coefficient, β is the overall bias (mean simulated value to mean observed value), and γ is the variance bias (ratio of simulated variance to observed variance).

The Critical Success Index (CSI) is calculated for events exceeding a threshold of 10 mm d^{-1} as:

$$\text{CSI} = \frac{H}{H + M + F + 10^{-9}}, \tag{B2}$$

where H is the number of hits (correctly predicted events), M is the number of misses (events missed by the product), and F is the number of false alarms (incorrectly predicted events). An epsilon value is added to prevent division by zero.

Peak bias at the 99.5th percentile (B_{peak}; %) is calculated as the percentage difference between the 99.5th percentile of the estimated and observed data:

$$B_{\text{peak}} = 100 \times \frac{P_{99.5} - O_{99.5}}{O_{99.5}}, \tag{B3}$$

where $P_{99.5}$ and $O_{99.5}$ are the 99.5th percentiles of the estimated and observed values, respectively.

Wet day bias ($B_{\text{wet days}}$; days) is calculated as the percentage difference in the number of wet days (days exceeding a 0.5 mm d^{-1} threshold) between simulated and observed data:

$$B_{\text{wet days}} = 365.25 \times \frac{P - O}{N}, \tag{B4}$$

where P and O are the number of wet days in the estimated and observed time series, respectively, and N is the total number of values.

Author contributions. XW: modeling, analysis, visualization, and writing. HEB: initial idea, conceptualization, modeling, analysis, writing, and project management. All coauthors contributed to writing, revising, and refining the manuscript.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. We sincerely acknowledge the developers of the datasets listed in Tables 1 and 2 for their efforts in creating and sharing these valuable resources. We gratefully acknowledge Prof. Mansour Almazroui (King Abdulaziz University, Jeddah, Saudi Arabia) for suggesting the name 'SaRa' for the product. This research was supported in part by funding from KAUST's Center of Excellence for Generative AI, under award number 5940.

References

Abbas, A., Yang, Y., Pan, M., Tramblay, Y., Shen, C., Ji, H., Gebrechorkos, S. H., Pappenberger, F., Pyo, J., Feng, D., Huffman, G., Nguyen, P. D., Massari, C., Brocca, L., and Beck, H. E.: Comprehensive Global Assessment of 23 Gridded Precipitation Datasets Across 16,295 Catchments Using Hydrological Modeling, Hydrology and Earth System Sciences Discussions, submitted, 2025.

Adhikari, A. and Behrangi, A.: Assessment of satellite precipitation products in relation with orographic enhancement over the western United States, Earth and Space Science, 9, e2021EA001 906, 2022.

Adler, R. F., Sapiano, M. R., Huffman, G. J., Wang, J.-J., Gu, G., Bolvin, D., Chiu, L., Schneider, U., Becker, A., and Nelkin, E.: The Global Precipitation Climatology Project (GPCP) monthly analysis (new version 2.3) and a review of 2017 global precipitation, Atmosphere, 9, 138, publisher: MDPI, 2018.

Al-Falahi, A. H., Saddique, N., Spank, U., Gebrechorkos, S. H., and Bernhofer, C.: Evaluation the performance of several gridded precipitation products over the highland region of yemen for water resources management, Remote Sensing, 12, 2984, publisher: MDPI, 2020.

Al-Ibrahim, A. A.: Excessive use of groundwater resources in Saudi Arabia: Impacts and policy options, Ambio, pp. 34–37, 1991.

Al Saud, M.: Assessment of flood hazard of Jeddah area 2009, Saudi Arabia, 2010.

Alharbi, R. S., Dao, V., Jimenez Arellano, C., and Nguyen, P.: Comprehensive Evaluation of Near-Real-Time Satellite-Based Precipitation: PDIR-Now over Saudi Arabia, Remote Sensing, 16, 703, https://doi.org/10.3390/rs16040703, number: 4 Publisher: Multidisciplinary Digital Publishing Institute, 2024.

Almazroui, M.: Calibration of TRMM rainfall climatology over Saudi Arabia during 1998–2009, Atmospheric Research, 99, 400–414, 2011.

Almazroui, M.: Rainfall Trends and Extremes in Saudi Arabia in Recent Decades, Atmosphere, 11, 964, https://doi.org/10.3390/atmos11090964, number: 9 Publisher: Multidisciplinary Digital Publishing Institute, 2020.

Almazroui, M., Islam, M. N., Saeed, S., Saeed, F., and Ismail, M.: Future changes in climate over the Arabian Peninsula based on CMIP6 multimodel simulations, Earth Systems and Environment, 4, 611–630, 2020.

Beck, H. E., Vergopolan, N., Pan, M., Levizzani, V., van Dijk, A. I. J. M., Weedon, G. P., Brocca, L., Pappenberger, F., Huffman, G. J., and Wood, E. F.: Global-scale evaluation of 22 precipitation datasets using gauge observations and hydrological modeling, Hydrology and Earth System Sciences, 21, 6201–6217, 2017.

Beck, H. E., Pan, M., Roy, T., Weedon, G. P., Pappenberger, F., van Dijk, A. I. J. M., Huffman, G. J., Adler, R. F., and Wood, E. F.: Daily evaluation of 26 precipitation datasets using Stage-IV gauge-radar data for the CONUS, Hydrology and Earth System Sciences, 23, 207–224, 2019a.

Beck, H. E., Wood, E. F., Pan, M., Fisher, C. K., Miralles, D. G., van Dijk, A. I. J. M., McVicar, T. R., and Adler, R. F.: MSWEP V2 global 3-hourly 0.1° precipitation: methodology and quantitative assessment, Bulletin of the American Meteorological Society, 100, 473–500, 2019b.

Beck, H. E., Westra, S., Tan, J., Pappenberger, F., Huffman, G. J., McVicar, T. R., Gründemann, G. J., Vergopolan, N., Fowler, H. J., Lewis, E., Verbist, K., and Wood, E. F.: PPDIST, global 0.1° daily and 3-hourly precipitation probability distribution climatologies for 1979–2018, Scientific Data, 7, https://doi.org/10.1038/s41597-020-00631-x, 2020.

Blenkinsop, S., Fowler, H. J., Barbero, R., Chan, S. C., Guerreiro, S. B., Kendon, E., Lenderink, G., Lewis, E., Li, X.-F., and Westra, S.: The INTENSE project: using observations and models to understand the past, present and future of sub-daily rainfall extremes, Advances in Science and Research, 15, 117–126, publisher: Copernicus Publications Göttingen, Germany, 2018.

515 Breiman, L.: Random forests, Machine Learning, 45, 5–32, 2001.

Brocca, L., Filippucci, P., Hahn, S., Ciabatta, L., Massari, C., Camici, S., Schüller, L., Bojkov, B., and Wagner, W.: SM2RAIN–ASCAT (2007–2018): global daily satellite rainfall data from ASCAT soil moisture observations, Earth System Science Data, 11, 1583–1601, https://doi.org/10.5194/essd-11-1583-2019, publisher: Copernicus GmbH, 2019.

Cannon, A. J., Sobie, S. R., and Murdock, T. Q.: Bias correction of GCM precipitation by quantile mapping: how well do methods preserve
520 changes in quantiles and extremes?, Journal of Climate, 28, 6938–6959, 2015.

Chen, M., Shi, W., Xie, P., Silva, V. B. S., Kousky, V. E., Wayne Higgins, R., and Janowiak, J. E.: Assessing objective techniques for gauge-based analyses of global daily precipitation, Journal of Geophysical Research: Atmospheres, 113, https://doi.org/10.1029/2007JD009132, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2007JD009132, 2008.

Chen, T. and Guestrin, C.: Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd acm sigkdd international conference on
525 knowledge discovery and data mining, pp. 785–794, 2016.

Ciabatta, L., Massari, C., Brocca, L., Gruber, A., Reimer, C., Hahn, S., Paulik, C., Dorigo, W., Kidd, R., and Wagner, W.: SM2RAIN-CCI: a new global long-term rainfall data set derived from ESA CCI soil moisture, Earth System Science Data, 10, 267–280, https://doi.org/10.5194/essd-10-267-2018, publisher: Copernicus GmbH, 2018.

Ciach, G. J.: Local Random Errors in Tipping-Bucket Rain Gauge Measurements, Journal of Atmospheric and Oceanic Technology, 20,
530 752–759, 2003.

Daly, C., Gibson, W. P., Taylor, G. H., Doggett, M. K., and Smith, J. I.: Observer bias in daily precipitation measurements at United States cooperative network stations, Bulletin of the American Meteorological Society, 88, 899–912, 2007.

Daly, C., Halbleib, M., Smith, J. I., Gibson, W. P., Doggett, M. K., Taylor, G. H., Curtis, J., and Pasteris, P. P.: Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States, International Journal of Climatology, 28,
535 2031–2064, https://doi.org/10.1002/joc.1688, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/joc.1688, 2008.

Danielson, J. J. and Gesch, D. B.: Global multi-resolution terrain elevation data 2010 (GMTED2010), Tech. Rep. 2331-1258, US Geological Survey, 2011.

Derin, Y., Anagnostou, E., Berne, A., Borga, M., Boudevillain, B., Buytaert, W., Chang, C.-H., Delrieu, G., Hong, Y., Hsu, Y. C., et al.: Multiregional satellite precipitation products evaluation over complex terrain, Journal of Hydrometeorology, 17, 1817–1836, 2016.

540 Dotse, S.-Q., Larbi, I., Limantol, A. M., and De Silva, L. C.: A review of the application of hybrid machine learning models to improve rainfall prediction, Modeling Earth Systems and Environment, 10, 19–44, https://doi.org/10.1007/s40808-023-01835-x, 2024.

Ebert, E. E., Janowiak, J. E., and Kidd, C.: Comparison of near-real-time precipitation estimates from satellite observations and numerical models, Bulletin of the American Meteorological Society, 88, 47–64, 2007.

El Kenawy, A. M. and McCabe, M. F.: A multi-decadal assessment of the performance of gauge- and model-based rainfall products over
545 Saudi Arabia: climatology, anomalies and trends, International Journal of Climatology, 36, 656–674, 2016.

El Kenawy, A. M., McCabe, M. F., Lopez-Moreno, J. I., Hathal, Y., Robaa, S. M., Al Budeiri, A. L., Jadoon, K. Z., Abouelmagd, A., Eddenjal, A., Domínguez-Castro, F., Trigo, R. M., and Vicente-Serrano, S. M.: Spatial assessment of the performance of multi-ple high-resolution satellite-based precipitation data sets over the Middle East, International Journal of Climatology, 39, 2522–2543, https://doi.org/10.1002/joc.5968, 2019.

550 Ensor, L. A. and Robeson, S. M.: Statistical characteristics of daily precipitation: Comparisons of gridded and point datasets, Journal of Applied Meteorology and Climatology, 47, 2468–2476, publisher: American Meteorological Society, 2008.

Fowler, H. J., Lenderink, G., Prein, A. F., Westra, S., Allan, R. P., Ban, N., Barbero, R., Berg, P., Blenkinsop, S., Do, H. X., Guerreiro, S., Haerter, J. O., Kendon, E. J., Lewis, E., Schaer, C., Sharma, A., Villarini, G., Wasko, C., and Zhang, X.: Anthropogenic intensification of short-duration rainfall extremes, Nature Reviews Earth & Environment, 2, 107–122, https://doi.org/10.1038/s43017-020-00128-6, 2021.

555 Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., Husak, G., Rowland, J., Harrison, L., Hoell, A., and Michaelsen, J.: The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes, Scientific Data, 2, 150 066, https://doi.org/10.1038/sdata.2015.66, 2015.

Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F.: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, Journal of hydrology, 377, 80–91, publisher: Elsevier, 2009.

560 Harris, I., Osborn, T. J., Jones, P., and Lister, D.: Version 4 of the CRU TS monthly high-resolution gridded multivariate climate dataset, Scientific Data, 7, https://doi.org/10.1038/s41597-020-0453-3, 2020.

Hasanean, H. and Almazroui, M.: Rainfall: Features and Variations over Saudi Arabia, A Review, Climate, 3, 578–626, 2015.

He, X., Chaney, N. W., Schleiss, M., and Sheffield, J.: Spatial downscaling of precipitation using adaptable random forests, Water resources research, 52, 8217–8237, publisher: Wiley Online Library, 2016.

565 Helmi, A. M. and Abdelhamed, M. S.: Evaluation of CMORPH, PERSIANN-CDR, CHIRPS V2.0, TMPA 3B42 V7, and GPM IMERG V6 Satellite Precipitation Datasets in Arabian Arid Regions, Water, 15, https://doi.org/10.3390/w15010092, 2023.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horanyi, A., noz Sabater, J. M., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., Chiara, G. D., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Holm, E., 570 Janiskova, M., Keeley, S., Laloyaux, P., Lopez, P., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J.-N.: The ERA5 global reanalysis, Quarterly Journal of the Royal Meteorological Society, https://doi.org/10.1002/qj.3803, 2020.

Hong, Y., Hsu, K.-L., Sorooshian, S., and Gao, X.: Precipitation Estimation from Remotely Sensed Imagery Using an Artificial Neural Network Cloud Classification System, Journal of Applied Meteorology and Climatology, 43, 1834–1853, https://doi.org/10.1175/JAM2173.1, publisher: American Meteorological Society Section: Journal of Applied Meteorology and Climatology, 2004.

575 Hrachowitz, M., Savenije, H., Blöschl, G., McDonnell, J. J., Sivapalan, M., Pomeroy, J., Arheimer, B., Blume, T., Clark, M., and Ehret, U.: A decade of Predictions in Ungauged Basins (PUB)—a review, Hydrological sciences journal, 58, 1198–1255, publisher: Taylor & Francis, 2013.

Huffman, G. J.: The transition in multi-satellite products from TRMM to GPM (TMPA to IMERG), 2019.

Huffman, G. J., Bolvin, D. T., Nelkin, E. J., Wolff, D. B., Adler, R. F., Gu, G., Hong, Y., Bowman, K. P., and Stocker, E. F.: The TRMM 580 Multisatellite Precipitation Analysis (TMPA): quasi-global, multiyear, combined-sensor precipitation estimates at fine scales, Journal of Hydrometeorology, 8, 38–55, 2007.

Huffman, G. J., Stocker, E. F., Bolvin, D. T., Nelkin, E. J., and Tan, J.: GPM IMERG final precipitation L3 half hourly 0.1 degree x 0.1 degree V06, Goddard Earth Sciences Data and Information Services Center (GES DISC): Greenbelt, MD, USA, 2019.

Huffman, G. J., Adler, R. F., Behrangi, A., Bolvin, D. T., Nelkin, E. J., Gu, G., and Ehsani, M. R.: The new version 3.2 global precip- 585 itation climatology project (GPCP) monthly and daily precipitation products, Journal of Climate, 36, 7635–7655, publisher: American Meteorological Society, 2023.

Hussein, E. A., Ghaziasgar, M., Thron, C., Vaccari, M., and Jafta, Y.: Rainfall prediction using machine learning models: literature survey, Artificial Intelligence for Data Science in Theory and Practice, pp. 75–108, publisher: Springer, 2022.

Intergovernmental Panel on Climate Change (IPCC): Climate Change 2021 – The Physical Science Basis: Working Group I Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge University Press, Cambridge, https://doi.org/10.1017/9781009157896, 2023.

Iturbide, M., Fernández, J., Gutiérrez, J. M., Bedia, J., Cimadevilla, E., Díez-Sierra, J., Manzanas, R., Casanueva, A., Baño-Medina, J., and Milovac, J.: Repository supporting the implementation of FAIR principles in the IPCC-WG1 Atlas, Zenodo, 2022, 2021.

Jazem Ghanim, A. A., Anjum, M. N., Alharbi, R. S., Aurangzaib, M., Zafar, U., Rehamn, A., Irfan, M., Rahman, S., Faraj Mursal, S. N., and Alyami, S.: Spatiotemporal evaluation of five satellite-based precipitation products under the arid environment of Saudi Arabia, AIP Advances, 14, publisher: AIP Publishing, 2024.

Joyce, R. J., Janowiak, J. E., Arkin, P. A., and Xie, P.: CMORPH: A Method that Produces Global Precipitation Estimates from Passive Microwave and Infrared Data at High Spatial and Temporal Resolution, Journal of Hydrometeorology, 5, 487–503, https://doi.org/10.1175/1525-7541(2004)005<0487:CAMTPG>2.0.CO;2, publisher: American Meteorological Society Section: Journal of Hydrometeorology, 2004.

Karger, D. N., Conrad, O., Böhner, J., Kawohl, T., Kreft, H., Soria-Auza, R. W., Zimmermann, N. E., Linder, H. P., and Kessler, M.: Climatologies at high resolution for the earth's land surface areas, Scientific Data, 4, 170 122, https://doi.org/10.1038/sdata.2017.122, number: 1 Publisher: Nature Publishing Group, 2017.

Kheimi, M. M. and Gutub, S.: Assessment of remotely-sensed precipitation products across the Saudi Arabia region, International Journal of Water Resources and Arid Environment, 4, 76–88, 2015.

Kidd, C., Becker, A., Huffman, G. J., Muller, C. L., Joe, P., Skofronick-Jackson, G., and Kirschbaum, D. B.: So, how much of the Earth's surface is covered by rain gauges?, Bulletin of the American Meteorological Society, 98, 69–78, 2017.

Kingdom of Saudi Arabia: Vision 2030: National Transformation Program Delivery Plan 2021–2025, https://www.vision2030.gov.sa/v2030/vrps/ntp/, 2021.

Kling, H., Fuchs, M., and Paulin, M.: Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios, Journal of hydrology, 424, 264–277, publisher: Elsevier, 2012.

Knoben, W. J., Freer, J. E., and Woods, R. A.: Inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores, Hydrology and Earth System Sciences, 23, 4323–4331, publisher: Copernicus GmbH, 2019.

Kochendorfer, J., Rasmussen, R., Wolff, M., Baker, B., Hall, M. E., Meyers, T., Landolt, S., Jachcik, A., Isaksen, K., and Brækkan, R.: The quantification and correction of wind-induced precipitation measurement errors, Hydrology and Earth System Sciences, 21, 1973–1989, publisher: Copernicus GmbH, 2017.

Kosaka, Y., Kobayashi, S., Harada, Y., Kobayashi, C., Naoe, H., Yoshimoto, K., Harada, M., Goto, N., Chiba, J., Miyaoka, K., Sekiguchi, R., Deushi, M., Kamahori, H., Nakaegawa, T., Tanaka, T., Tokuhiro, T., Sato, Y., Matsushita, Y., and Onogi, K.: The JRA-3Q Reanalysis, Journal of the Meteorological Society of Japan. Ser. II, 102, https://doi.org/10.2151/jmsj.2024-004, 2024.

Kubota, T., Aonashi, K., Ushio, T., Shige, S., Takayabu, Y. N., Kachi, M., Arai, Y., Tashima, T., Masaki, T., and Kawamoto, N.: Global Satellite Mapping of Precipitation (GSMaP) products in the GPM era, Satellite Precipitation Measurement: Volume 1, pp. 355–373, publisher: Springer, 2020.

Lewis, E., Fowler, H., Alexander, L., Dunn, R., McClean, F., Barbero, R., Guerreiro, S., Li, X.-F., and Blenkinsop, S.: GSDR: A Global Sub-Daily Rainfall Dataset, Journal of Climate, 32, 4715–4729, https://doi.org/10.1175/JCLI-D-18-0143.1, publisher: American Meteorological Society Section: Journal of Climate, 2019.

Lin, J., Qian, T., Bechtold, P., Grell, G., Zhang, G. J., Zhu, P., Freitas, S. R., Barnes, H., and Han, J.: Atmospheric convection, Atmosphere-Ocean, 60, 422–476, 2022.

Lin, Y. and Mitchell, K. E.: 1.2 the NCEP stage II/IV hourly precipitation analyses: Development and applications, in: Proceedings of the 19th Conference Hydrology, American Meteorological Society, San Diego, CA, USA, vol. 10, Citeseer, https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=22e13910ffd515119f42c7c63edb71c239a67c58, 2005.

Mahmoud, M. T., Al-Zahrani, M. A., and Sharif, H. O.: Assessment of global precipitation measurement satellite products over Saudi Arabia, Journal of Hydrology, 559, 1–12, https://doi.org/10.1016/j.jhydrol.2018.02.015, 2018.

Massari, C., Brocca, L., Pellarin, T., Abramowitz, G., Filippucci, P., Ciabatta, L., Maggioni, V., Kerr, Y., and Fernandez Prieto, D.: A daily 25 km short-latency rainfall product for data-scarce regions based on the integration of the Global Precipitation Measurement mission rainfall and multiple-satellite soil moisture products, Hydrology and Earth System Sciences, 24, 2687–2710, https://doi.org/10.5194/hess-24-2687-2020, publisher: Copernicus GmbH, 2020.

Menne, M. J., Durre, I., Vose, R. S., Gleason, B. E., and Houston, T. G.: An Overview of the Global Historical Climatology Network-Daily Database, Journal of Atmospheric and Oceanic Technology, 29, 897–910, https://doi.org/10.1175/JTECH-D-11-00103.1, publisher: American Meteorological Society Section: Journal of Atmospheric and Oceanic Technology, 2012.

Munir, S., Kamil, S., Habeebullah, T. M., Zaidi, S., Alhajji, Z., and Siddiqui, M. H.: Drought variabilities in Saudi Arabia: Unveiling Spatiotemporal Trends through Observations and Projections, Earth Systems and Environment, pp. 1–25, 2025.

New, M., Hulme, M., and Jones, P.: Representing Twentieth-Century Space-Time Climate Variability. Part I: Development of a 1961–90 Mean Monthly Terrestrial Climatology, Journal of Climate, 12, 829–856, https://doi.org/10.1175/1520-0442(1999)012<0829:RTCSTC>2.0.CO;2, 1999.

Nguyen, P., Ombadi, M., Sorooshian, S., Hsu, K., AghaKouchak, A., Braithwaite, D., Ashouri, H., and Thorstensen, A. R.: The PERSIANN family of global satellite precipitation data: a review and evaluation of products, Hydrology and Earth System Sciences, 22, 5801–5816, https://doi.org/10.5194/hess-22-5801-2018, publisher: Copernicus GmbH, 2018.

Nguyen, P., Ombadi, M., Gorooh, V. A., Shearer, E. J., Sadeghi, M., Sorooshian, S., Hsu, K., Bolvin, D., and Ralph, M. F.: PERSIANN Dynamic Infrared–Rain Rate (PDIR-Now): A Near-Real-Time, Quasi-Global Satellite Precipitation Dataset, Journal of Hydrometeorology, 21, 2893–2906, https://doi.org/10.1175/JHM-D-20-0177.1, publisher: American Meteorological Society Section: Journal of Hydrometeorology, 2020.

O, S. and Foelsche, U.: Assessment of spatial uncertainty of heavy rainfall at catchment scale using a dense gauge network, Hydrology and Earth System Sciences, 23, 2863–2875, https://doi.org/10.5194/hess-23-2863-2019, publisher: Copernicus GmbH, 2019.

Osborn, T. and Hulme, M.: Development of a relationship between station and grid-box rainday frequencies for climate model evaluation, Journal of Climate, 10, 1885–1908, publisher: American Meteorological Society, 1997.

Othman, A., El-Saoud, W. A., Habeebullah, T., Shaaban, F., and Abotalib, A. Z.: Risk assessment of flash flood and soil erosion impacts on electrical infrastructures in overcrowded mountainous urban areas under climate change, Reliability Engineering & System Safety, 236, 109 302, https://doi.org/10.1016/j.ress.2023.109302, 2023.

Overeem, A., van den Besselaar, E., van der Schrier, G., Meirink, J. F., van der Plas, E., and Leijnse, H.: EURADCLIM: the European climatological high-resolution gauge-adjusted radar precipitation dataset, Earth System Science Data, 15, 1441–1464, https://doi.org/10.5194/essd-15-1441-2023, publisher: Copernicus GmbH, 2023.

Papacharalampous, G., Tyralis, H., Doulamis, A., and Doulamis, N.: Comparison of Machine Learning Algorithms for Merging Gridded Satellite and Earth-Observed Precipitation Data, Water, 15, 634, https://doi.org/10.3390/w15040634, number: 4 Publisher: Multidisciplinary Digital Publishing Institute, 2023.

665 Patlakas, P., Stathopoulos, C., Flocas, H., Bartsotas, N. S., and Kallos, G.: Precipitation Climatology for the Arid Region of the Arabian Peninsula — Variability, Trends and Extremes, Climate, 9, https://doi.org/10.3390/cli9070103, 2021.

Peters, K., Hohenegger, C., and Klocke, D.: Different representation of mesoscale convective systems in convection-permitting and convection-parameterizing NWP models and its implications for large-scale forecast evolution, Atmosphere, 10, 503, 2019.

Prakash, S.: Performance assessment of CHIRPS, MSWEP, SM2RAIN-CCI, and TMPA precipitation products across India, Journal of
670 Hydrology, 571, 50–59, https://doi.org/10.1016/j.jhydrol.2019.01.036, 2019.

Sadeghi, M., Nguyen, P., Naeini, M. R., Hsu, K., Braithwaite, D., and Sorooshian, S.: PERSIANN-CCS-CDR, a 3-hourly 0.04° global precipitation climate data record for heavy precipitation studies, Scientific Data, 8, 157, https://doi.org/10.1038/s41597-021-00940-9, publisher: Nature Publishing Group, 2021.

Sevruk, B., Ondrás, M., and Chvíla, B.: The WMO precipitation measurement intercomparisons, Atmospheric Research, 92, 376–380, 2009.

675 Shen, Z., Yong, B., Gourley, J. J., Qi, W., Lu, D., Liu, J., Ren, L., Hong, Y., and Zhang, J.: Recent global performance of the Climate Hazards group Infrared Precipitation (CHIRP) with Stations (CHIRPS), Journal of Hydrology, 591, 125 284, https://doi.org/10.1016/j.jhydrol.2020.125284, 2020.

Sivapalan, M., Takeuchi, K., Franks, S., Gupta, V., Karambiri, H., Lakshmi, V., Liang, X., McDonnell, J., Mendiondo, E. M., and O'connell, P.: IAHS Decade on Predictions in Ungauged Basins (PUB), 2003–2012: Shaping an exciting future for the hydrological sciences, Hydro-
680 logical sciences journal, 48, 857–880, publisher: Taylor & Francis, 2003.

Smith, A., Lott, N., and Vose, R.: The Integrated Surface Database: Recent Developments and Partnerships, Bulletin of the American Meteorological Society, 92, 704–708, https://www.jstor.org/stable/26218543, publisher: American Meteorological Society, 2011.

Sultan, M., Sturchio, N. C., Alsefry, S., Emil, M. K., Ahmed, M., Abdelmohsen, K., AbuAbdullah, M. M., Yan, E., Save, H., Alharbi, T., Othman, A., and Chouinard, K.: Assessment of age, origin, and sustainability of fossil aquifers: A geochemical and remote sensing-based
685 approach, Journal of Hydrology, 576, 325–341, https://doi.org/https://doi.org/10.1016/j.jhydrol.2019.06.017, 2019.

Sun, Q., Miao, C., Duan, Q., Ashouri, H., Sorooshian, S., and Hsu, K.-L.: A review of global precipitation datasets: data sources, estimation, and intercomparisons, Reviews of Geophysics, 56, 79–107, 2018.

Tabari, H. and Willems, P.: Seasonally varying footprint of climate change on precipitation in the Middle East, Scientific Reports, 8, https://doi.org/10.1038/s41598-018-22795-8, 2018.

690 Tang, G., Behrangi, A., Long, D., Li, C., and Hong, Y.: Accounting for spatiotemporal errors of gauges: A critical step to evaluate gridded precipitation products, Journal of Hydrology, 559, 294–306, https://doi.org/10.1016/j.jhydrol.2018.02.057, 2018.

Ting, Y.-S.: Why Machine Learning Models Systematically Underestimate Extreme Values, arXiv preprint arXiv:2412.05806, 2024.

Trabucco, A. and Zomer, R. J.: Global aridity index and potential evapotranspiration (ET0) climate database v2, CGIAR Consort Spat Inf, 10, m9, 2018.

695 Villarini, G., Mandapaka, P. V., Krajewski, W. F., and Moore, R. J.: Rainfall and sampling uncertainties: A rain gauge perspective, Journal of Geophysical Research: Atmospheres, 113, 2008.

Wang, Y., You, Y., and Kulie, M.: Global Virga Precipitation Distribution Derived From Three Spaceborne Radars and Its Contribution to the False Radiometer Precipitation Detection, Geophysical Research Letters, 45, 4446–4455, https://doi.org/10.1029/2018GL077891, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2018GL077891, 2018.

700 Xie, P., Joyce, R., Wu, S., Yoo, S.-H., Yarosh, Y., Sun, F., and Lin, R.: Reprocessed, Bias-Corrected CMORPH Global High-Resolution Precipitation Estimates from 1998, Journal of Hydrometeorology, 18, 1617–1641, https://doi.org/10.1175/JHM-D-16-0168.1, publisher: American Meteorological Society Section: Journal of Hydrometeorology, 2017.

Xu, Y., Tang, G., Li, L., and Wan, W.: Multi-source precipitation estimation using machine learning: Clarification and benchmarking, Journal of Hydrology, 635, 131 195, publisher: Elsevier, 2024.

705 Yamamoto, M. K., Shige, S., Yu, C.-K., and Cheng, L.-W.: Further improvement of the heavy orographic rainfall retrievals in the GSMaP algorithm for microwave radiometers, Journal of Applied Meteorology and Climatology, 56, 2607–2619, 2017.

Yang, S., Jones, P. D., Jiang, H., and Zhou, Z.: Development of a near-real-time global in situ daily precipitation dataset for 0000–0000 UTC, International Journal of Climatology, 40, 2795–2810, 2020.

Yano, J.-I., Ziemiański, M. Z., Cullen, M., Termonia, P., Onvlee, J., Bengtsson, L., Carrassi, A., Davy, R., Deluca, A., Gray, S. L., et al.:
710 Scientific challenges of convective-scale numerical weather prediction, Bulletin of the American Meteorological Society, 99, 699–710, 2018.

Yates, E., Anquetin, S., Ducrocq, V., Creutin, J.-D., Ricard, D., and Chancibault, K.: Point and areal validation of forecast precipitation fields, Meteorological Applications, 13, 1–20, https://doi.org/10.1017/S1350482705001921, 2006.

Youssef, A. M., Sefry, S. A., Pradhan, B., and Alfadail, E. A.: Analysis on causes of flash flood in Jeddah city (Kingdom of Saudi
715 Arabia) of 2009 and 2011 using multi-sensor remote sensing data and GIS, Geomatics, Natural Hazards and Risk, 7, 1018–1042, https://doi.org/10.1080/19475705.2015.1012750, publisher: Taylor & Francis _eprint: https://doi.org/10.1080/19475705.2015.1012750, 2016.

Table 1. Overview of quasi- and fully-global *P* products used in this study. Abbreviations: P=precipitation; T=temperature; S=satellite; G=gauge; Re=reanalysis; A=analysis; NRT=near-real-time; Pred=used as predictor to generate SaRa; and Eval=included in the performance assessment. The column for spatial coverage denotes "Global" for complete global coverage including ocean regions, and "Land" for coverage limited to terrestrial areas. Version information unavailable for most products.

Data	Version	Variables	Data Source	Resolution		Coverage		Time Latency	References/URL	Utilization
				Temporal	Spatial	Temporal	Spatial			
IMERG-E	V07	P	S	30 min	0.1°	2000–NRT	60° N/S	~4 hours	Huffman et al. (2019)	Eval, Pred
IMERG-L	V07	P	S	30 min	0.1°	2000–NRT	60° N/S	~12 hours	Huffman et al. (2019)	Eval, Pred
IMERG-F	V07	P	S, G	30 min	0.1°	2000–NRT	60° N/S	~3 months	Huffman et al. (2019)	Eval
GSMaP-NRT	V8	P	S	hourly	0.1°	2000–NRT	60° N/S	~4 hours	Kubota et al. (2020)	Eval, Pred
GSMaP-MVK	V8	P	S	hourly	0.1°	2000–NRT	60° N/S	~3 days	Kubota et al. (2020)	Eval, Pred
ERA5	/	P	Re	hourly	0.25°	1940–NRT	Global	~5 days	Hersbach et al. 2020	Eval, Pred
ERA5	/	T	Re	hourly	0.25°	1940–NRT	Global	~5 days	Hersbach et al. (2020),	Pred
GDAS	/	P	A	hourly	0.25°	2001–NRT	Global	~3–6 hours	[1]	Eval, Pred
GDAS	/	T	A	hourly	0.25°	2001–NRT	Global	~3–6 hours	[1]	Pred
PDIR-Now	/	P	S	hourly	0.04°	2000–NRT	60° N/S	~100 minutes	Nguyen et al. (2020)	Eval, Pred
PERSIANN-CCS-CDR	/	P	S, G	3-hourly	0.04°	1983–2021	60° N/S	/	Sadeghi et al. (2021)	Eval, Pred
JRA-3Q	/	P	Re	3-hourly	~40 km	1947–NRT	Global	~20 days	Kosaka et al. (2024)	Eval
CMORPH-RAW	/	P	S	30 min	~8 km	2019–NRT	60°N/S	~4 hours	Joyce et al. (2004)	Eval
CMORPH-RT	/	P	S	30 min	~8 km	2019–NRT	60°N/S	~4 hours	Xie et al. (2017)	Eval
PERSIANN-CCS	/	P	S	Hourly	0.04°	2003–NRT	60° N/S	~90 minutes	Hong et al. (2004)	Eval
CPC Unified	/	P	G	Daily	0.5°	1979–NRT	Land	~1 day	Chen et al. (2008)	Eval
SM2RAIN-CCI	/	P	S	Daily	0.25°	1998–2015	Land	/	Ciabatta et al. (2018)	Eval
SM2RAIN-ASCAT	/	P	S	Daily	0.1°	2007–2021	Land	/	Brocca et al. (2019)	Eval
SM2RAIN-GPM	/	P	S	Daily	0.25°	2007–2018	Land	/	Massari et al. (2020)	Eval
CHIRP	V2	P	S, Re, A	Daily	0.05°	1981–NRT	Land, 50° N/S	~6 days	Funk et al. (2015)	Eval
CHIRPS	V2	P	S, G, Re, A	Daily	0.05°	1981–NRT	Land, 50° N/S	2 weeks	Funk et al. (2015)	Eval
MSWEP	V2.8	P	S, G, Re, A	3-hourly	0.1°	1979–NRT	Global	~3 hours	Beck et al. (2019b)	Eval
SaRa	V1	P	S, G, Re, A	Hourly	0.1°	1979–NRT	Global	~100 minutes	This paper	Eval

[1] https://www.ncei.noaa.gov/products/weather-climate-models/global-data-assimilation

Table 2. Overview of the static predictors used in the ML models to generate our new P product, SaRa.

Name (units)	Data Source(s)	Description
AI ($-$)	Mean annual P from CHELSA V2.1 (1-km resolution; Karger et al., 2017) and PET from Trabucco and Zomer (2018, 1 km) for land and ERA5 (0.25°) for ocean	Aridity index (AI) calculated as ratio of P to PET
Pmean ($-$)	CHELSA V2.1 (1-km resolution; Karger et al., 2017)	Mean annual P
ETH (m)	Global Multi-resolution Terrain Elevation Data (GMTED) 2010 (Danielson and Gesch, 2011)	Effective Terrain Height (ETH) calculated following Daly et al. (2008)
Lat (°)	/	Latitude
Lon (°)	/	Longitude
AbsLat (°)	/	Absolute latitude

Table 3. The dynamic predictors incorporated in the different ML model stacks.

	ERA5 P	ERA5 T	GDAS P	GDAS T	IMERG-L V07	GSMaP-MVK V8	P-CCS-CDR	PDIR-Now
01	✓	✓	✗	✗	✓	✓	✗	✗
02	✓	✓	✗	✗	✓	✗	✗	✗
03	✓	✓	✗	✗	✗	✓	✓	✗
04	✓	✓	✗	✗	✗	✗	✗	✗
05	✓	✓	✗	✗	✓	✓	✗	✗
06	✓	✗	✓	✓	✓	✓	✗	✗
07	✗	✓	✓	✓	✓	✗	✗	✗
08	✓	✓	✓	✓	✗	✓	✗	✗
09	✓	✗	✓	✓	✓	✓	✗	✗
10	✗	✗	✓	✓	✓	✗	✗	✗
11	✗	✗	✗	✗	✗	✓	✗	✗
12	✗	✗	✗	✗	✓	✗	✗	✓
13	✗	✗	✗	✗	✓	✓	✗	✗
14	✗	✗	✓	✓	✗	✗	✗	✗
15	✗	✗	✓	✓	✗	✗	✗	✓
16	✓	✓	✓	✓	✗	✗	✗	✗
17	✗	✗	✓	✓	✗	✗	✗	✗
18	✗	✗	✗	✗	✗	✗	✗	✓

Table 4. The performance of the ML model stacks underlying our new P product SaRa (models_01–18) and other state-of-the-art P products sorted in descending order of median KGE. The values represent medians calculated over all randomly selected independent evaluation gauges in Saudi Arabia. Note that since the different P products span different temporal periods, the specific evaluation data used for the evaluation differs between P products. The unit for B_{peak} is %, and the unit for $B_{wet\,days}$ is number of days. N_{obs} indicates the number of stations used to assess each product.

	KGE	r_{dly}	β	$\lvert\beta-1\rvert$	γ	$\lvert\gamma-1\rvert$	r_{mon}	B_{peak}	$\lvert B_{peak}\rvert$	$B_{wet\,days}$	$\lvert B_{wet\,days}\rvert$	$CSI_{10\,mm}$	N_{obs}
model_06	0.43	0.55	1.03	0.22	0.94	0.15	0.75	0.83	24.77	0.00	2.89	0.22	119
model_08	0.41	0.53	0.94	0.27	0.95	0.13	0.75	-5.68	28.49	0.39	2.48	0.20	119
model_03	0.39	0.49	0.94	0.24	0.98	0.13	0.69	-16.56	26.77	1.38	2.57	0.20	119
model_07	0.39	0.51	1.01	0.25	0.94	0.14	0.73	3.64	27.67	1.06	3.21	0.21	119
model_09	0.39	0.51	0.88	0.28	0.99	0.15	0.74	-14.60	30.84	-0.39	3.17	0.22	119
model_11	0.37	0.53	1.02	0.28	0.93	0.15	0.73	-3.34	28.28	0.00	3.12	0.22	119
model_01	0.36	0.50	1.03	0.24	0.96	0.12	0.71	-11.17	27.01	1.42	2.58	0.21	119
model_02	0.36	0.46	0.93	0.27	1.02	0.12	0.69	-19.14	31.26	0.43	2.08	0.20	119
model_10	0.34	0.48	1.02	0.28	0.96	0.13	0.72	-10.42	33.33	0.34	2.62	0.19	119
model_12	0.32	0.46	1.01	0.25	0.89	0.13	0.64	-7.59	26.20	3.44	4.35	0.18	119
model_14	0.32	0.44	1.06	0.21	0.84	0.18	0.61	-10.88	22.57	3.71	4.31	0.17	119
model_15	0.31	0.45	0.93	0.30	0.90	0.17	0.72	-15.25	30.62	1.68	3.28	0.16	119
model_13	0.29	0.41	1.04	0.25	0.87	0.15	0.57	-3.85	24.42	2.42	3.35	0.17	119
IMERG-F V07	0.29	0.50	1.16	0.27	0.76	0.24	0.73	-8.96	28.78	10.92	10.92	0.18	119
model_16	0.28	0.40	1.04	0.32	0.95	0.17	0.66	-15.42	38.45	1.53	3.47	0.14	119
GDAS	0.24	0.42	1.09	0.32	0.72	0.28	0.66	-20.57	35.23	12.80	12.80	0.12	119
model_17	0.23	0.38	0.89	0.37	0.94	0.18	0.62	-9.37	40.42	0.44	3.75	0.11	119
CMORPH-RT	0.21	0.39	0.69	0.40	0.97	0.21	0.56	-44.55	48.38	1.66	4.18	0.10	119
model_04	0.21	0.38	0.75	0.35	1.05	0.18	0.62	-37.17	44.83	0.20	2.81	0.12	118
ERA5	0.21	0.36	0.99	0.24	0.75	0.28	0.61	-30.18	39.06	11.82	11.82	0.11	119
MSWEP V2.8	0.20	0.42	1.03	0.31	0.64	0.38	0.67	-28.65	44.58	18.77	18.77	0.11	118
SM2RAIN-GPM	0.18	0.40	0.79	0.44	0.66	0.36	0.60	-47.16	53.43	10.75	11.09	0.00	84
CHIRPS V2	0.17	0.30	0.96	0.28	0.72	0.28	0.55	-27.82	38.12	6.25	6.42	0.09	118
model_18	0.13	0.29	0.90	0.37	0.89	0.22	0.45	-24.39	33.54	4.26	6.98	0.09	119
JRA-3Q	0.12	0.29	1.16	0.41	0.72	0.29	0.59	-18.02	38.92	9.79	9.84	0.10	119
PDIR-Now	0.11	0.32	1.34	0.47	0.80	0.26	0.47	-5.89	29.44	16.30	16.30	0.10	119
model_05	0.03	0.31	0.55	0.53	1.15	0.25	0.51	-60.18	64.33	-2.55	3.85	0.09	119
CPC-Unified	-0.03	0.21	0.53	0.54	1.13	0.23	0.51	-47.80	51.57	-1.84	4.06	0.04	113
CHIRP V2	-0.07	0.23	1.10	0.27	0.35	0.65	0.39	-56.68	57.45	44.36	44.36	0.00	119
GSMaP-MVK V8	-0.13	0.46	1.98	0.98	0.81	0.23	0.62	44.70	45.54	15.74	15.74	0.19	119
PERSIANN-CCS-CDR	-0.18	0.31	1.84	0.84	0.78	0.23	0.49	62.49	62.83	11.36	11.65	0.11	118
CMORPH-RAW	-0.19	0.44	2.06	1.06	0.82	0.21	0.57	77.54	77.54	10.80	10.95	0.18	119
SM2RAIN-ASCAT	-0.20	0.19	1.27	0.34	0.27	0.73	0.31	-67.16	67.41	49.68	49.68	0.00	117
SM2RAIN-CCI	-0.24	0.18	0.42	0.71	0.33	0.70	0.38	-83.05	85.41	5.25	20.16	0.00	65
IMERG-L V07	-0.39	0.45	2.22	1.22	0.75	0.26	0.61	72.91	72.91	14.75	14.75	0.16	119
IMERG-E V07	-0.44	0.42	2.26	1.26	0.71	0.31	0.59	67.20	67.20	19.79	19.79	0.16	119
PERSIANN-CCS	-0.72	0.31	2.54	1.54	0.67	0.33	0.39	114.09	114.09	15.57	15.57	0.09	119

Table 5. Spatial Spearman rank correlation coefficients between rain gauge attributes (related to location, climate, and topography; see Table 2) and performance scores (KGE, r_{dly}, β, etc.; see Section 2.6) for six key P products and ML model stacks. The correlations were computed using the independent evaluation gauges in Saudi Arabia ($n = 119$).

Model	Predictor	KGE	r_{dly}	β	$|\beta-1|$	γ	$|\gamma-1|$	r_{mon}	B_{peak}	$|B_{\mathrm{peak}}|$	$B_{\mathrm{wet\,days}}$	$|B_{\mathrm{wet\,days}}|$	$\mathrm{CSI}_{10\,\mathrm{mm}}$
	Lat	0.09	0.39	0.26	0.03	0.04	0.17	0.17	0.31	0.09	−0.30	−0.38	0.40
	Lon	0.05	0.17	−0.01	0.08	0.06	−0.04	0.32	0.07	0.01	0.05	0.42	−0.04
model_01	Pmean	−0.13	−0.42	−0.32	0.00	0.00	−0.08	−0.24	−0.29	−0.02	0.22	0.48	−0.39
	AI	0.14	0.43	0.28	−0.01	0.00	0.09	0.24	0.28	0.03	−0.24	−0.47	0.40
	ETH	−0.13	−0.42	−0.13	−0.08	−0.13	0.03	−0.33	−0.21	−0.07	0.34	0.36	−0.28
	Lat	0.13	0.24	0.03	−0.03	0.45	−0.01	0.07	0.15	0.09	−0.44	−0.39	0.04
	Lon	−0.06	0.06	0.22	0.20	−0.10	−0.05	0.12	0.19	0.12	0.03	0.34	0.02
model_06	Pmean	−0.05	−0.20	0.01	−0.02	−0.36	−0.04	−0.13	−0.12	−0.13	0.32	0.48	0.03
	AI	0.08	0.24	0.00	0.01	0.40	0.02	0.12	0.15	0.12	−0.35	−0.48	−0.01
	ETH	0.08	−0.16	−0.12	−0.22	−0.17	0.01	−0.19	−0.29	−0.33	0.19	0.31	0.08
	Lat	0.47	0.68	0.34	−0.05	0.19	−0.24	0.53	0.50	−0.21	−0.71	−0.72	0.67
	Lon	−0.22	−0.01	−0.12	0.20	0.00	0.08	0.08	−0.11	0.09	0.26	0.26	−0.23
ERA5	Pmean	−0.44	−0.64	−0.33	0.11	0.00	0.01	−0.48	−0.42	0.18	0.69	0.71	−0.57
	AI	0.47	0.67	0.32	−0.08	0.06	−0.06	0.51	0.42	−0.20	−0.72	−0.73	0.58
	ETH	−0.22	−0.50	−0.15	0.01	−0.02	−0.02	−0.41	−0.28	0.01	0.51	0.53	−0.36
	Lat	0.25	0.40	0.12	−0.05	0.43	−0.37	0.24	0.47	−0.12	−0.81	−0.81	0.19
	Lon	−0.08	0.05	−0.20	0.14	0.15	0.05	0.13	−0.23	0.12	0.16	0.18	−0.22
MSWEP V2.8	Pmean	−0.17	−0.29	−0.21	−0.02	−0.35	0.29	−0.17	−0.60	0.17	0.79	0.80	−0.18
	AI	0.21	0.29	0.16	−0.01	0.37	−0.29	0.17	0.57	−0.15	−0.82	−0.83	0.19
	ETH	−0.03	−0.22	−0.12	−0.19	−0.36	0.21	−0.20	−0.48	0.12	0.60	0.60	−0.01
	Lat	−0.41	0.24	0.42	0.42	−0.14	0.16	−0.08	0.48	0.48	−0.31	−0.31	−0.03
	Lon	0.05	0.18	−0.02	−0.02	0.02	0.04	0.40	−0.11	−0.13	0.43	0.44	0.08
IMERG−L V07	Pmean	0.52	−0.31	−0.52	−0.52	0.27	−0.29	0.04	−0.61	−0.61	0.28	0.28	0.08
	AI	−0.53	0.29	0.53	0.54	−0.25	0.27	−0.07	0.62	0.62	−0.27	−0.27	−0.08
	ETH	0.39	−0.37	−0.41	−0.40	0.33	−0.37	−0.15	−0.47	−0.46	0.08	0.08	0.09
	Lat	−0.21	0.39	0.23	0.23	−0.27	0.12	0.20	0.20	0.19	−0.48	−0.48	0.21
	Lon	0.32	0.14	−0.27	−0.26	0.27	−0.20	0.38	−0.19	−0.15	0.08	0.08	0.25
GSMaP-MVK V8	Pmean	0.36	−0.33	−0.38	−0.37	0.40	−0.23	−0.20	−0.32	−0.30	0.48	0.48	−0.05
	AI	−0.33	0.35	0.35	0.35	−0.37	0.21	0.20	0.32	0.30	−0.49	−0.49	0.08
	ETH	0.07	−0.34	−0.12	−0.12	0.16	−0.03	−0.35	−0.17	−0.15	0.54	0.54	−0.06

Table 6. Predictor importances (%) for each submodel of model_01. Predictors are ranked in descending order of importance for Submodel 1. $\hat{y}_{\text{Submodel 2}}$ represents the 3-hourly P output of Submodel 2.

Predictor	Submodel 1	Submodel 2	Submodel 3	Submodel 4
ERA5 P	41.85	14.96	–	13.47
IMERG-L V07	35.59	59.60	–	43.77
GSMaP-MVK V8	5.48	9.72	–	28.85
Lat	2.81	3.21	–	2.64
Lon	2.85	1.95	–	2.00
Pmean	2.67	1.92	0.46	1.63
ERA5 T	2.42	2.46	0.41	2.83
AbsLat	2.36	2.40	0.99	1.87
AI	2.00	1.85	–	1.36
ETH	1.96	1.94	–	1.59
$\hat{y}_{\text{Submodel 2}}$	–	–	98.14	–

Table A1. Hyperparameters for the XGBoost models.

Hyperparameter	Value
n_estimators	100
max_depth	12
min_child_weight	5
colsample_bytree	0.7
gamma	2
reg_alpha	0.5
reg_lambda	0.5
learning_rate	0.2

Table A2. Hyperparameters for the RF models.

Hyperparameter	Value
n_estimators	100
max_depth	15
min_samples_split	5
min_samples_leaf	5
max_features	0.7