Journal of the Civil Engineering Forum, January 2025, 11(1):97-108
DOI 10.22146/jcef.13242
Available Online at https://jurnal.ugm.ac.id/v3/jcef/issue/archive

Assessing Basin's Dynamic Hydrological Characteristics Using Statistical Analysis On Rainfall – River Discharge Observation Data

Steven Reinaldo Rusli*, Theo Senjaya
Civil Engineering Department, Universitas Katolik Parahyangan, Bandung, INDONESIA
*Corresponding author: steven.reinaldo@unpar.ac.id

SUBMITTED 10 May 2024 **REVISED** 31 July 2024 **ACCEPTED** 06 August 2024

ABSTRACT Hydrological studies often rely on physical-based modelling approaches to simulate water cycles. However, such an approach requires extensive basin physical data inputs, including features, attributes, and properties that are quantifiable, which often are lacking in data-scarce areas. Therefore, this study explores an alternative viewpoint by using simple statistical analysis to assess the dynamic basin's hydrological characteristics. We collate and divide the rainfall and discharge observation data in the Upper Citarum River basin into three periods: period 1 (2000–2005), period 2 (2000–2010), and period 3 (2000–2015). After defining baseflow separation, we quantify the basin's baseflow using simple statistical analysis. The 5-year average of the baseflow fluctuations (33.15 m^3s^{-1}, 12.88 m^3s^{-1}, and 27.59 m^3s^{-1} during each period) agrees with previous studies' physical-based results. The subsequent frequency analysis indicates a trend of increasing rainfall, although it is not followed by the trend in the river discharge variable. Due to the stochastic nature of extreme events occurrence and available data length, we evaluate the dynamic basin's runoff generation using quasi-synthetic rainfall instead of conventional design storm to equalize the stimuli (rainfall) in evaluating the target system (basin's hydrological characteristics). Under identical sets of forcing input, the quasi-synthetic river discharge consistently increases in each period in both the median (15.39% and 25.34%) and extreme (21.86% and 29.46%) values. The results reveal the basin's evolving hydrological responses, which are mostly influenced by anthropogenic factors. This simple statistical approach enables the evaluation of basin characteristics' dynamics in data-limited areas, bypassing extensive data collection and random event occurrences while still providing consistent results.

KEYWORDS Hydrological characteristics; Statistic-based assessment; Baseflow; Runoff generation; Upper Citarum River basin

1 INTRODUCTION

In many hydrological studies, physical-based models are often relied upon for basin assessment due to their nature of real-world representation (Al-Areeq et al., 2021; Gelete et al., 2023). Such an approach is established on the concept of translating basin features and properties into model parameters, leading to more intuitive model setup. With the recent advancements in remote-sensing technologies and open sciences, many global datasets on basin physical characteristics become easily accessible, positively contributing to the growing the popularity of physical-based hydrological models' applications. Some of the online datasets, for example, are Digital Elevation Model (DEM) (Yamazaki et al., 2017*a*; Abrams et al., 2020), soil type (Shangguan et al., 2014; Poggio et al., 2021), land use (Potapov et al., 2022), river network (Munier and Decharme, 2022; Yan et al., 2022), and more.

However, physical-based modelling also presents several drawbacks. The heterogeneous nature of most basin's attributes often propagates to demands of spatially high- resolution models, further lead to sub-grid scale integration and problems of dimensionality in parameter calibration (Beven, 1989). Even in situations where the the study area is relatively homogeneous, it still requires an extensive amount of basin physical data. Physical-based models are also dominantly regulated by the model structure (Muhammad et al., 2019; van Kempen et al., 2021), therefore selecting an appropriate model for each case is crucially important.

In response to these limitations, parsimonious modelling approach (Paniconi and Putti, 2015; Newman et al., 2017; Rusli et al., 2023) become favored in data-limited areas. Simple models capable of representing hydrological processes are preferable over complex models with insufficient justification in their setup and parameterization. In the midst of balancing between model representation and model complexity, a statistical approach emerges. In general, statistic-based analysis requires fewer data inputs yet produces still reliable performances (Wang et al., 2021; Sun et al., 2023). In data-scarce areas with limited data availability, statistic-based analysis has the potential to unravel the hydrological processes of the associated study area.

In this study, we apply the simple yet reliable statistical approach to unravel and quantify the dynamic hydrological characteristics in the test basin of the Upper Citarum River basin. Using exclusively the rain-

fall and river discharge observation data without any prior information on the basin's physical attributes, we aim to accurately quantify the river baseflow estimates and assess the runoff generation process within the study area. Through this approach, we achieve all the aforementioned objectives, bypassing the tedious basin's physical data collection and pre-processing, while keeping the analysis results consistent and satisfactory. We expect the proposed method to be applicable to other study areas, assisting in understanding hydrological processes, particularly from the perspective of the river baseflow estimates and runoff generation responses.

2 MATERIALS AND METHODS

2.1 Study area

The study area is the Upper Citarum River basin (UCRB), located in the West Java Province, Indonesia (Figure 1). Its area includes the Bandung City and several other surrounding districts, covering a total area of approximately 1,817 km^2. As the name suggests, the main river flowing through the outlet is the upper part of the Citarum River, as shown in Figure 1.

Topographically, the UCRB exhibits a high spatial variation in terrain elevation. As depicted in Figure 1, the the elevation range of the basin spans from 645 meters above sea level at the outlet to a high 2,571 meters above sea level at its peak, estimated from the Multi-Error-Removed Improved-Terrain Digital Elevation Model (MERIT-DEM) dataset (Yamazaki et al., 2017*b*). From a hydrological perspective, the steep-sloped area is prone to high runoff generation, while the vast plain terrain at the middle of the basin serves as the receiver of the surface flow. The slope of the river naturally mirrors the terrain, resulting in slower flow velocity at the downstream part of the basin.

2.2 Rainfall and river discharge observation data

Within the UCRB, eight rainfall stations' data are available, covering the period between 2000 and 2015. To address the rainfall's spatial variability, each station's spatial weight is calculated using the Polygon Thiessen method (see Table 1). All coverage areas are in accordance with the standard proposed by the World Meteorological Organization (WMO), taking into account the coverage and topographical distributions of the study area. Nonetheless, Java Island is already the hydrologically- densest measured island in Indonesia. This increases the degree of confidence in using the UCRB as the designated study area. Figure 2 presents the boxplot of each rainfall station's measurement and the areal rainfall estimates.

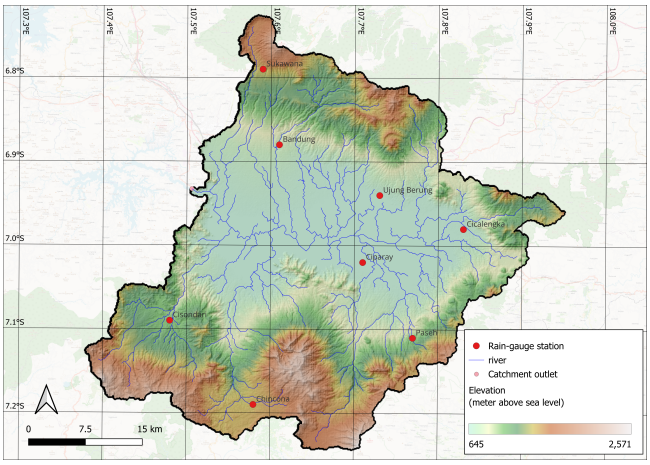

Figure 1 The general overview of the Upper Citarum river basin. Eight rain-gauge stations are identified within the basin boundary, while an Automatic Water Level Recorder (AWLR) for measuring daily river discharge is installed at the catchment outlet. The range of terrain elevation is represented by the color bar.

Table 1. Rainfall stations' weight over the Upper Citarum River basin area, approximated using Polygon Thiessen

Station	Area (km^2)	Weight (%)
Sukawana	97.84	5.38
Cisondari	300.96	16.56
Ujungberung	232.61	12.80
Cicalengka	204.86	11.27
Ciparay	259.81	14.30
Paseh	166.85	9.18
Chincona	258.79	164.24
Bandung	295.55	16.26

The UCRB is also equipped with an Automatic Water Level Recorder (AWLR). The temporal coverage of the used river discharge observation data at the basin outlet is consistent with the available rainfall data. Some measurements, however, are contentious, particularly in 2004 and 2011, where the observed river discharge data is relatively lower in comparison to the areal rainfall estimates. These data are later excluded in further calculations. Figure 3 displays the time-series discharge observation, accompanied by the areal rainfall values, to demonstrate the consistency between the stimuli (rainfall) and response (runoff) variables.

In evaluating the basin characteristics' dynamic, in particular the river baseflow and runoff generation process, we split the data duration into three periods: period 1 between 2000 and 2005, period 2 between 2006 and 2010, and period 3 between 2011 and 2015. These intervals are selected based on the land use/land cover (LULC) changes detected in previous studies (Agaton et al., 2016; Kuntoro et al., 2018). While these studies used decadal time steps for examining the changes in the basin's characteristics, it is evident that the changes are significant in each time step. Given the notable figures, we divide the analysis periods into three phases to produce more detailed results and analysis.

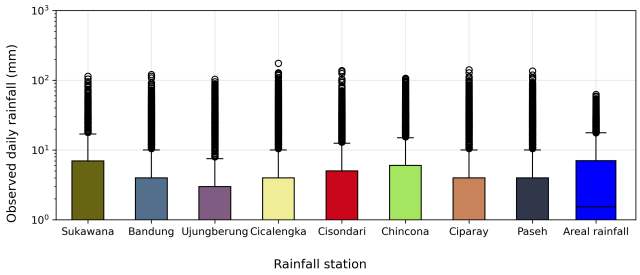

Figure 2 Comparison between the rainfall measurements in each station and the areal rainfall estimates' distribution. The narrower range of the areal rainfall estimates relative to the station's data reflects the non linearity of (extreme) event occurrences.

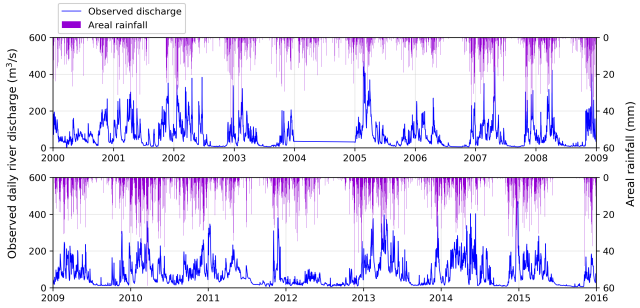

Figure 3 Time-series plot of the areal rainfall (right axis) and river discharge observation data (left axis).

2.3 Baseflow separation

Several physical-based methods are commonly used to estimate river baseflow, starting from parameter filtering (Ladson et al., 2013; Zhou et al., 2021), baseflow index (Oktavia et al., 2022; Taufik and Annisa', 2022) to other physical process-based approaches (Furey and Gupta, 2001; Duncan, 2019; Sun et al., 2021). In this study, we propose a simple approach by focusing on the basic definition of river baseflow. Essentially, river baseflow is the portion of streamflow that, for the most part, is sustained, even during little to no precipitation period. It is fed and maintained by the surrounding groundwater flow, seeping through the riverbeds and banks, provided that the groundwater head remains higher than the riverbed elevation. Thus, our first criterion for separating baseflow from river discharge data is to select observed discharge data from days with 'little to no rainfall'.

However, 'little to no rainfall' terminology is qualitative. As mentioned above, there are eight rainfall stations in the UCRB used in this study. Even when all but one rainfall station reports no rainfall, the areal rainfall estimates would still report a rainfall event should merely one station experience and report one measurement. This non-linearity of event occurrences is visible in Figure 2, resulting in a higher median of the areal rainfall estimates compared to measurements from individual rainfall stations. As a result, there are only a few instances of areal rainfall being recorded as non-rainy days (32.20%). In context, the average number of non-rainy days across all rainfall stations is calculated at 61.21%, consistent with the climatic characteristics of the study area. Therefore, we apply the 'little to no rainfall' threshold condition at 1 mm. With this benchmark, the count of non-rainy days increases to 45.89%.

On the other side of the above situation, on days with no rainfall, there are possibilities of high river discharge observed values if that particular data were recorded following direct day(s) of heavy rainfall. Due to its moderate size, the rainfall falling upstream of the UCRB may take more than a day to reach the basin's outlet, delaying the river discharge observation data by one timestep compared to the measured rainfall. These two distinct events – the 'little to no rainfall' and the consecutive rainfall events (CRE) – have to be considered in baseflow estimation. Therefore, in this study, we collate all the river discharge measured on the non-rainy days and use their median value for each period to avoid extreme values and establish our simple statistic-based baseflow estimates.

2.4 Frequency analysis

Frequency analysis is a statistical method that relates the magnitude of probable extreme events to the frequency of their occurrences, typically expressed in return periods. A higher return period corresponds to a lower probability and frequency of extreme event occurrence, along with a higher magnitude of the analyzed variable. In this study, the frequency analysis is applied to both the rainfall data and the river discharge observation data under 2, 5, 10, 25, 50, and 100-year return periods. The formula for such analyses, resulting in design storm and design flood, is written as follows:

$$x_T = \bar{x} + k.s \tag{1}$$

With x_T representing the design storm (mm) or design flood (m^3s^{-1}) under the selected return period, $\bar{x}$ and s representing the mean and the standard deviation of the annual maximum daily rainfall data (mm) or annual maximum daily discharge observation data (m^3s^{-1}), respectively, and k representing the frequency factor which depends on the probability distribution function (PDF) (dimensionless) and the selected return period.

Within the frequency analysis, there are numerous PDFs that can be applied to the formula. In this study, it is performed using the Generalized Extreme Value (GEV) distribution, which has been proven to perform well in many other studies, be it applied to rainfall data (Yoon et al., 2013; Ginting and Putuhena, 2017) or flood discharge data (Jiang and Kang, 2019; Samantaray and Sahoo, 2020). The formula for the GEV cumulative dis-

tribution function (CDF) is as follows:

$$F(x) = e^{\left\{-\left[1\frac{k(x-\xi)}{\alpha}\right]^{1/k}\right\}} \text{for } k \neq 0 \quad (2)$$

With ξ representing the location parameter, α the scale parameter, and *k* the shape parameter. For each *k* value, the GEV distribution corresponds to certain extreme value distribution types, with the moments for the GEV distribution described comprehensively in (Smith, 1965).

Typically, applying frequency analysis incorporates the inclusion of the whole available data. In this study, however, we also aim to assess the trends in the rainfall and the river discharge variable in the UCRB. Given the minimum data length of 10 years (Tunas and Oka, 2020) for hydrology-related frequency analysis and the available data of 16 years, we opt to assimilate the data progressively. The frequency analysis is iterated in 4 phases, using the data from 2000 to 2009, from 2000 to 2011, from 2000 to 2013, and from 2000 to 2015. To address the non-linearity issue on the extreme rainfall events occurrence – similar to the baseflow estimates analysis but on the other side of the 'extreme' – we first conduct the frequency analysis on each rainfall station before computing the areal rainfall estimates.

2.5 Statistical regression line fitting

After dealing with the low flow part of the dataset, we move our focus to the higher rainfall and river discharge section of the dataset, whose aerial rainfall estimates is higher than 1 mm. Using the identical period division as in the baseflow estimates section (period 1, period 2, and period 3), we compute the correlation trend line between the rainfall and river discharge using simple linear regression (least square method) function:

$$\hat{y} = a + b.x \quad (3)$$

With $\hat{y}$ representing the predicted discharge (m^3s^{-1}), *a* representing the y-intercept of the regression line (m^3s^{-1}), *b* representing the slope of the regression line ($m^3s^{-1}mm^{-1}$), and *x* representing the rainfall variable (mm). The value of *a* and *b* is calculated as follows:

$$b = \frac{\Sigma(xy) - \frac{\Sigma x \Sigma y}{n}}{\Sigma(x^2) - \frac{(\Sigma x)^2}{n}} \quad (4)$$

$$a = \frac{\Sigma y}{n} - b.\frac{\Sigma x}{n} \quad (5)$$

With *y* representing the dependent variable of river discharge (m^3s^{-1}) and *x* representing the independent variable of rainfall (mm). We evaluate the regression linear using correlation coefficient *r*, calculated using the following formula:

Table 2. Correlation coefficient (*r*) evaluation criteria

r value	Indication
0 < *r* < 0.3	Weak positive linear relationship
0.3 < *r* < 0.7	Moderate positive linear relationship
0.7 < *r* < 1.0	Strong positive linear relationship

$$r = \frac{\Sigma(x_i - \frac{\Sigma x}{n})(y_i - \frac{\Sigma y}{n})}{\sqrt{\Sigma(x_i - \frac{\Sigma x}{n})^2 \Sigma(y_i - \frac{\Sigma y}{n})^2}} \quad (6)$$

The resulting r values are evaluated based on the range proposed in (Ratner, 2009), presented in Table 2. When the application of the least square method to the rainfall and discharge correlation yields non-weak linear relationships, we use the trendline equation to project the design storm into another design flood approximation.

2.6 Quasi-synthetic rainfall and river discharge

To assess the basin's dynamic characteristics, we quantify the changes in the basin's response to a certain set of stimuli over the three designed periods. The stimuli refer to the forcing input, regulated entirely by rainfall. The basin's characteristics embody the target system, while the basin's response is constituted by river discharge. However, alongside the dynamics of the target system, the forcing input does also fluctuate. The assessment of the basin's dynamic characteristics under fluctuating rainfall does involve systematic bias, as the basin's response is influenced by both the forcing input and the target system's attributes and properties. To ensure a robust evaluation, the target system is forced by equal stimuli. This warrants that changes in the basin's responses are attributed solely to the basin's properties and not the forcing input. Therefore, to standardize (and equalize) the forcing input, we create sets of rainfall dataset, named the 'quasi-synthetic rainfall' to approximate sets of quasi-synthetic river discharge.

The term 'quasi-synthetic' is used as these datasets primarily consist of hypothetical data. However, it does not mean that they are not fully imaginary. The quasi-synthetic rainfall, for example, is derived based on the results of frequency analysis. A series of rainfall events ranging from zero (non-rainy days) to the values of the design storm is established identically for each analyzed return period. This ensures the rainfall heights to be aligned with the actual measurement data, while still representing the changes in the input variable through different return periods, hence 'quasi-synthetic rainfall'. These quasi-synthetic rainfalls are then forced into the target system. The resulting basin's responses, termed 'quasi-synthetic river discharge', are approximated using the least-square re-

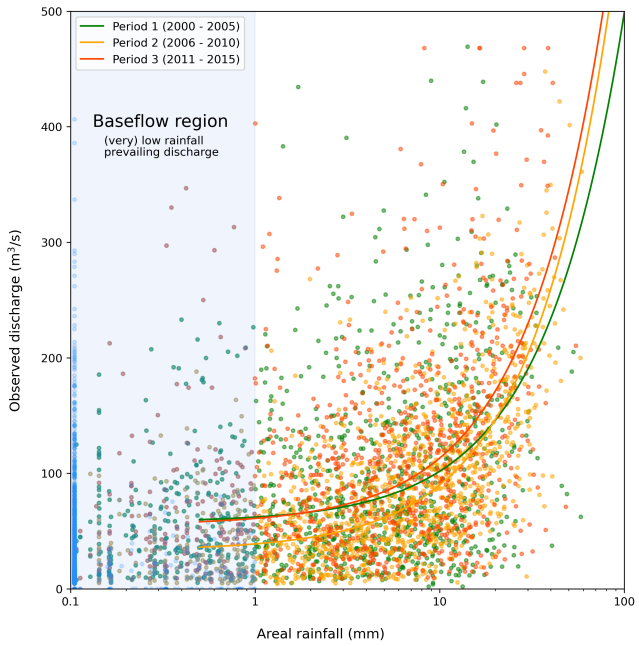

Figure 4 Scatter plot of the areal rainfall estimates and river discharge observation data. The blue region on the left separates the 'little to no rainfall' area used for baseflow estimates. The rainfall and river discharge correlation data are categorized into three periods: period 1 (green), period 2 (yellow), and period 3 (red). The trend lines of each category are also presented, showing increasing slope over the periods.

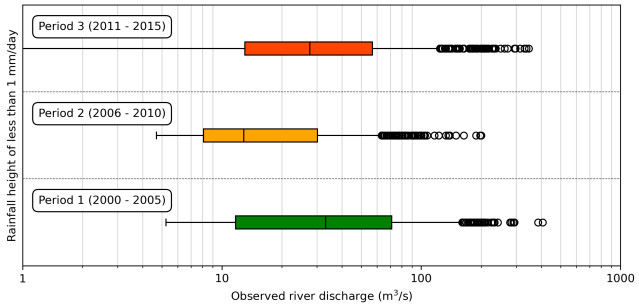

Figure 5 Distribution of river discharge baseflow estimates for each period in focus. The median values of the boxplots (33.15 m^3s^{-1} for period 1, 12.88 m^3s^{-1} for period 2, and 27.59 m^3s^{-1} for period 3) represent the 5- years average baseflow estimates, showing the fluctuating nature of the Citarum River.

gression equation described in the previous section for each period. As the quasi-synthetic river discharges are calculated based upon different sets of target system's trendline, the analysis is expected to produce different runoff generation processes for each period, which is further juxtaposed with the information on the LULC changes in the study area.

3 RESULTS

3.1 Baseflow and high-flow separation

Figure 4 separates the 'baseflow' (left) and the 'runoff generation' (right) areas. The presentation facilitates an intuitive correlation between variables; the *x*-axis is the areal rainfall (independent variable) in mm, and the *y*-axis is the observed discharge (dependent variable) in m^3s^{-1}. To accommodate the wide range and non-uniform distribution of rainfall data, the x-axis is plotted on a log scale. To prevent losses of many data points, we set the value of rainfall during non-rainy days to 0.1 mm as the log-scale plot cannot display zero-valued data points. While this adjustment affects a large amount of data, the compromise in data quality is minimal since 0.1 mm of rainfall is practically non-existent, especially when compared to the extreme rainfall events of higher return periods. The baseflow region in Figure 4 also shows a large variation in the river discharge observation data, ranging from 4.7 m^3s^{-1} to 406.4 m^3s^{-1}, despite the approximately similar estimated areal rainfall values. This is due to the CRE mentioned in Section 2.3. While the 'current non-rainy day' is included in the baseflow region section of Figure 4, the preceding rainy days are not. In practice, while the non-rainy days intuitively lead to low river discharge, prior hydrological events may regulate and influence the river discharge in subsequent time steps. This explains the presence of several data points where the river discharge is relatively high even during periods of low areal rainfall.

On the right side of Figure 4, the areal rainfall estimates and river discharge observation data pairings are plotted and categorized by period 1, period 2, and period 3. Then, we calculate and plot three regression lines, also for each period. To support intuitive visualization, both the dots and the regression lines are displayed with the same color scheme. The three regression lines appear parabolic despite having linear equations due to the semi-log plotting nature of Figure 4. All regression lines indicate a moderate to strong relationship between data pairings, with the *r* values of 0.40, 0.74, and 0.54 for period 1, period 2, and period 3, respectively. A trend is observed from the regression lines, too – the slope of each line increases sequentially across the periods: 4.41, 5.66, and 5.80 for period 1, period 2, and period 3 respectively.

3.2 Baseflow estimate

Figure 5 displays the horizontal boxplots of the data pairings in the baseflow region (Figure 4), categorized by different periods. Due to the wide range of discharge values under approximately similar low areal rainfall estimates, we approximate the river baseflow as the median of the data pairings. Figure 5 also features an inverse axis compared to Figure 4: the *x*-axis represents the river discharge observation data in m^3s^{-1}, while the *y*-axis depicts the areal rainfall estimates in mm. This adjustment provides a better visual, as the river discharge observation data have a significantly wider range in comparison to that of the areal rainfall estimates. The estimated 5-year average river base-

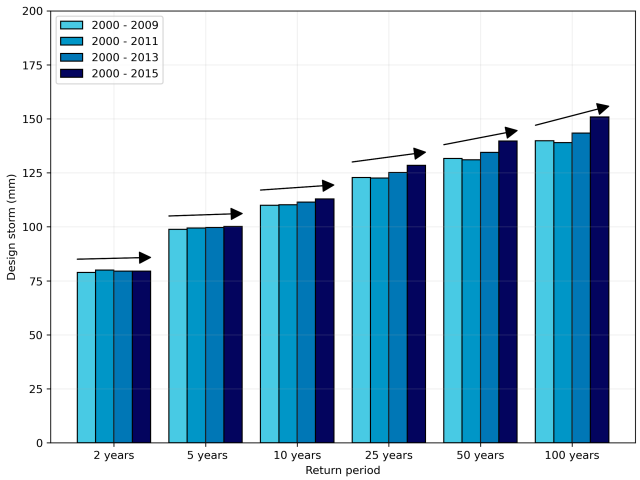

Figure 6 Results of progressive design storm under various return periods. The arrows above the bar charts represent the slope of changes in design storm values, considering the latest rainfall trend.

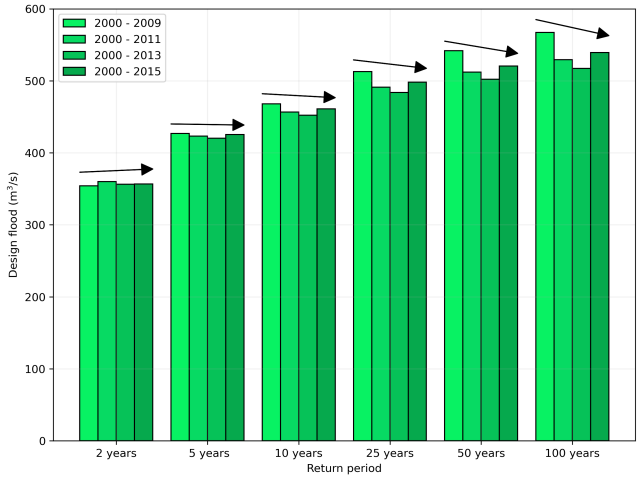

Figure 7 Results of progressive design flood under various return periods. The arrows above the bar charts represent the slope of changes in design flood values, considering the latest discharge trend.

flow estimates for each period (1, 2, and 3) are 33.15 m^3s^{-1}, 12.88 m^3s^{-1}, and 27.59 m^3s^{-1}, respectively. Our findings are either consistent or only slightly overestimate the results from previous studies, discussed in Section 4.3. Our approach to estimating river baseflow is proven to be capable of circumventing the laborious physical data collection and preprocessing required for physical-based baseflow separation analysis.

3.3 Progressive design storm and design flood

Following the baseflow analysis, Figures 6 and 7 depict the progressive frequency analysis results for the areal rainfall estimates and river discharge observation data, respectively. As outlined in Section 2.4, the frequency analysis is conducted progressively, resulting in four bar charts for each return period. Interestingly, Figure 6 and Figure 7 reveal two contrasting trends. In the frequency analysis of the rainfall variable (Figure 6), incorporating more data into the frequency analysis leads to higher design storm values for every return period, indicated by the arrows' direction above the bar charts. However, the values in the second bar chart have a uniquely diverging direction relative to the other changes for each return period, elaborated further in Section 4.4. It is also noted that the higher the return period is, the higher the gap is among the results of the progressive frequency analysis. Similarly, this is also discussed in Section 4.4.

On the contrary to the frequency analysis results for the design storm, the design flood decreases as the return period increases, with the exception of the two-year return period values (Figure 7). The magnitude also follows another trend: the decreases in the design flood are smaller under the lower return periods and are larger under the higher return periods. This is related to the length of the assimilated data and the variation of the annual peak discharge, discussed comprehensively in Section 4.4. Additionally, this decreasing trend is observed only until the third phase of the progressive frequency analysis, involving river discharge observation data from 2000 to 2013. When the river discharge observation data from 2014 and 2015 are included, the design flood increases again to a considerable extent, especially for the higher return periods. This U-turn drift is further discussed in Section 4.4.

3.4 Quasi-synthetic areal rainfall and river discharge

As outlined in Section 2.6, assessing the dynamics of the basin's characteristics is only meaningful under an equalized forcing input. Figure 6 reveals that the assumption of equal forcing is far from accurate among different periods. This is evidenced by the consistently changing trends of design storms resulting from the progressive frequency analysis. Therefore, relying solely on river discharge observation data in evaluating the basin's response may not yield robust conclusions, as forcing inputs play an important role in the river discharge generation, obviously alongside the basin's characteristic itself. Thus, equalized quasi-synthetic areal rainfall for each period is created using the range between zero and the frequency analysis results for each return period.

Aligned with the research aim of conducting statistic-based analysis rather than physical-based hydrological modelling, the quasi-synthetic rainfall is forced to the regression line equation between the areal rainfall estimates and the river discharge instead of a structured hydrological model. In computing the trendline using the least-square method described in Section 2.5, the areal rainfall estimates serve as the independent variable, while the river discharge as the dependent variable. The computation results of the quasi-synthetic river discharge are displayed in Figure 8, showing vary-

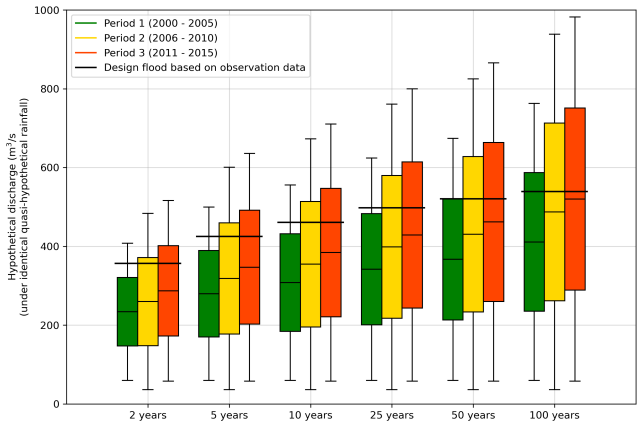

Figure 8 The quasi-synthetic river discharge derived from quasi-synthetic rainfall. The different color schemes represent the three periods constituted by unique regression line equations. The design flood calculated from the frequency analysis on the actual river discharge observation data is also displayed alongside the quasi-synthetic river discharge.

ing outcomes for each period. A clear trend emerges, the quasi-synthetic river discharge increases between every period with different degrees of the change. A higher return period leads to a higher increase between each period. Notably, the increment from period 1 to period 2 consistently exceeds that from period 2 to period 3, a phenomenon further examined and linked to changes in the basin's LULC, discussed in Section 4.4.

The design flood derived from the frequency analysis of the river discharge observation data under each return period is also plotted in Figure 8 (black lines), exhibiting values close to the upper quartile of the quasi-synthetic river discharge. However, as the return period increases, the disparity between these two variables becomes more pronounced. This discrepancy is attributable to inherent uncertainties and limitations, elaborated further in Section 4.1, as well as the non-linearity of extreme event occurrences, discussed in Section 4.2.

We also observe that the lower values of the quasi- synthetic river discharge mirror the trend of the baseflow estimates for each period. In the baseflow estimates depicted in Figure 5, it is estimated that the baseflow in the second period has the lowest values, while those in the first and third periods are more comparable. The lower ends of the quasi-synthetic river discharge boxplots in Figure 8 precisely reflect this trend, demonstrating the consistency of the proposed statistic-based analysis.

4 DISCUSSIONS

4.1 Limitations and uncertainties involved

In doing this research, we acknowledge that physical-based hydrological modeling offers greater intuitiveness and a more direct link with actual field conditions. Conversely, the statistic-based method circumvents the laborious tasks associated with data collection and pre-processing, and requires lower computational resources. Clearly, both methods have their own strengths and limitations, and each inevitably possesses inherent uncertainties, including in our own proposed method.

In our proposed approach for estimating baseflow, we encounter a broad range of river discharge observation data. While we suggest employing median values as an approximation for the baseflow, it's important to note that this assumption may not be universally applicable across all basins. Factors such as soil moisture and condition, topography, river bathymetry, and others play significant roles in baseflow generation. Therefore, in different regions, using other statistical attributes might yield more accurate results. This also highlights the potential for developing semi-physical methods, which combine less resource-intensive statistical computations with parsimonious physical-based analyses.

In addition to baseflow estimates, our proposed statistic- based approach encompasses the analysis of high-flow regions. This is achieved through statistical regression line fitting using the least-square method. The resulting trendline is represented by a linear equation, which in itself constitutes a considerable assumption. For future studies, exploring alternative regression methods commonly used in hydrology, such as polynomial (Ostertagová, 2012), fuzzy (Bardossy et al., 1990), or Bayesian regression (Reis et al., 2020), could yield new insights into statistic-based hydrological assessment. There is also concern about using daily-based data instead of event-based data, which occur in a shorter time scale in the UCRB. The annual maximum daily rainfall, although measured daily, actually occurs in hours; the same with the flood discharge. Future research could very well benefit from another perspective and analysis using event-based data.

The progressive frequency analysis on the rainfall and river discharge data demonstrates a contrasting trend: the design storm is predicted to increase when the latest observation data is incorporated, while the design flood decreases. This is an interesting tendency as, intuitively, increasing storm should lead to higher river discharge. There are several reasons behind this phenomenon. Firstly, it is related to the dataset's time step of daily interval. The used river discharge dataset in this study is 'averaged' on a daily basis instead of measured on instantaneous event-based river discharge. This averaging process reduces the estimated peak discharge, causing an underestimation of the actual flood event. Secondly, the frequency analysis is conducted based on only one-day measurement (the annual maximum daily flood discharge). Therefore, consecutive high discharge is not taken into account. In several observed years, while the annual maximum daily river dis-

charge is lower than others, its total outflow volume is actually higher. This indicates the difference in flood characteristics (short-high peak vs long-consistent stable peak) that is not taken into consideration in annual maximum daily data-based frequency analysis. Last but not least, the result of the frequency analysis is also influenced by sample size. Using (only) 16 years of annual maximum daily data generates a low sample size. Therefore, one or two diverging flood characteristics would sway the overall frequency analysis results, especially on the high return period.

Our approach to calculating quasi-synthetic river discharge also does not consider CRE (consecutive rainfall events). The regression lines are determined solely based on same-day data pairings; thus, any preceding rainfall events are not considered and may lead to an underestimation of river discharge estimates. Physically, CRE could potentially increase the 'current day' river discharge (Wu et al., 2009; Zheng et al., 2020). Moreover, it is projected to occur more frequently in the future, too (Du et al., 2022). In our study, potential effects stemming from this issue are mitigated by comparing the quasi-synthetic river discharge to the design flood. Fortunately, Figure 8 indicates that underestimation does not occur, as the design floods are calculated closer to the upper quartile of the quasi-synthetic river discharge values.

4.2 Non-linearity of extreme events occurrences

Non-linearity issues could occur both in the low and the high values of rainfall events. In the case of low rainfall events, to account for non-rainy days in the study area, all rainfall stations must record zero rainfall on the same day. However, this scenario is often not met. Therefore, as suggested in Section 2.3, we classify any areal rainfall estimates of less than 1 mm per day as non-rainy days.

A similar issue arises at the other end of the rainfall spectrum with high extreme rainfall events. Calculating quasi-synthetic areal rainfall estimates involves two steps: spatial analysis and frequency analysis. Frequency analysis is relatively straightforward – a series of annual maximum daily rainfall is collected and extrapolated. Spatial analysis, however, integrates the individual rainfall station's measurement to the basin's areal rainfall estimates. The issue of non-linearity occurs because an annual maximum daily rainfall measured in one station does not necessarily lead to one in other rainfall stations, which could potentially lead to an underestimation of areal design storm. To tackle this issue, we employ the frequency analysis for each rainfall station dataset before the spatial analysis. On the contrary, this approach may lead to overestimation as it assumes the annual maximum daily rainfall to occur on the same day at all stations. Addressing this issue in future studies may involve incorporating areal reduction factors into the analysis (Pietersen et al., 2015; Pavlovic et al., 2016; Mineo et al., 2019).

4.3 The dynamics of baseflow estimates

The variation among baseflow estimates in numerous studies is inevitable. For instance, in the study by (Salim et al., 2019), the UCRB's baseflow was calculated based on the water balance proportion. It was approximated that 30.29% of rainfall influx generated average baseflow of 29.94 m^3s^{-1} in 2015. This aligns very closely with our findings of the 5-years average baseflow of 27.59 m^3s^{-1} between 2011 and 2015. Similarly, our estimates also corroborated with values proposed by (Sudradjat et al., 2020).

Conversely, analyzing the period spanning from 2009 to 2018, (Fadhil et al., 2021) estimated baseflow values ranging from 4.3 m^3s^{-1} to 38.1 m^3s^{-1}. Although this may appear to align with our findings initially, their calculated average baseflow of 12.3 m^3s^{-1} falls below our estimates. The overestimation is also echoed by (Sebayang et al., 2022), with a baseflow estimates of 7 m^3s^{-1}, representing the lowest estimates thus far. These discrepancies underscore the need for further studies to expand our limited understanding of hydrological processes in the UCRB and possibly to account for the fluctuating nature of river baseflow annually. The temporal coverage of studies on baseflow estimates in the UCRB does vary, making year-by-year evaluations challenging. Moving forward, conducting baseflow estimate studies at higher temporal resolutions could prove beneficial.

4.4 The dynamics of runoff generation

Both Figure 6 and Figure 7 illustrate a global trend of increasing rainfall and decreasing river discharge. However, locally, there are contrasting trends, i.e. the second phase of the progressive frequency analysis on rainfall data and the fourth phase on the river discharge data. These discrepancies occur due to the fluctuating hydrological variables in the UCRB. For instance, 2011 was relatively dry in terms of annual maximum daily rainfall, ranking as the second driest year between 2000 and 2015. The inclusion of one low annual maximum daily rainfall value in the second phase of the progressive frequency analysis, with only 12 data points, significantly influences the results, especially in higher return periods. A similar phenomenon occurs in the frequency analysis of river discharge observation data, but in the opposite direction. The second highest river discharge record in the UCRB of 468.1 m^3s^{-1} occurred in 2014. Incorporating this specific data point into the last phase of the progressive frequency analysis substantially increases the resulting design flood. Previ-

ous studies have suggested that longer historical data records lead to more accurate frequency analysis (Kobierska et al., 2018; Tunas and Oka, 2020), in particular to those with high return periods. Therefore, in this study, the highest considered return period is 100 years, as conducting frequency analysis based on 16 years of data would not result in accurate estimates for higher return periods.

As assessing the dynamic's basin characteristics is the main aim of this study, it is crucial to observe the increment of the river discharge estimates among each period. Under identical forcings, the river discharge increases in period 1, period 2, and period 3. Since the forcings are equalized, the heightened river discharge must result solely from the evolving basin's response to generate runoff. Numerous studies have suggested that runoff generation is influenced by two primary driving factors: natural and anthropogenic factors (Rezaei et al., 2019; Hou et al., 2020; Wang et al., 2022). Natural factors relate to the basin's intrinsic attributes, such as vegetation cover, soil saturation, and topography, while anthropogenic factors pertain to urban development, LULC changes, impervious surfaces, etc. While natural factors change slowly, anthropogenic factors are more dynamic. The quasi-synthetic river discharge indicates a change in the basin's response, which is supported by previous studies (Agaton et al., 2016; **?**; Fadhil et al., 2021). Increasing runoff could be attributable to the forest area reduction of 35% reported by (Agaton et al., 2016). On the other hand, the moderately developed area coverage also increases from 10.5% to 23.1% from 1985 to 2015 (Kuntoro et al., 2018). Our results are consistently in agreement even when compared in more detail timesteps with long-term LULC change studies. The diverging gaps between period 1 to period 2 and period 2 to period 3 in Figure 8 agree with the results reported by (Yulianto et al., 2020), as the LULC changes between period 1 and period 2 are more significant than those between period 2 and period 3, as indicated by the higher percentage of primary forest area loss.

5 CONCLUSION

In this study, a statistic-based approach for assessing basin's dynamic hydrological characteristics is developed with the aim of bypassing the tedious data collection and pre-processing typically encountered in physical-based hydrological modelling frameworks. Using the rainfall and river discharge observation data in the UCRB between 2000 and 2015, we succeeded in quantifying the 5-years average river baseflow: 33.15 m^3s^{-1}, 12.88 m^3s^{-1}, and 27.59 m^3s^{-1}, which aligns with estimates from most previous studies. To quantify the changes in the basin's response (river discharge) to rainfall, we equalized the forcing for each period using a quasi-synthetic rainfall, developed from the progressive frequency analysis on areal rainfall estimates. The results indicate that under standardized rainfall under various return periods, the basin's response is evolving due to anthropogenic factors influencing the runoff generation processes. The quasi-synthetic river discharge consistently increases in each period, in both the median (15.39% and 25.34%) and extreme (21.86% and 29.46%) values. These findings are corroborated by previous studies focusing on LULC development in the UCRB. The proposed methods and findings are expected to be applicable in areas with limited basin physical data and are able to circumvents the necessity for laborious pre-simulation work, while assisting in understanding the hydrological processes in various regions of application.

DISCLAIMER

The authors declare no conflict of interest.

REFERENCES

Abrams, M., Crippen, R. and Fujisada, H. (2020), 'Aster global digital elevation model (gdem) and aster global water body dataset (astwbd)', *Remote Sensing* **12**(7), 1156.
URL: *https://doi.org/10.3390/rs12071156*

Agaton, M., Setiawan, Y. and Effendi, H. (2016), 'Land use/land cover change detection in an urban watershed: A case study of upper citarum watershed, west java province, indonesia', *Procedia Environmental Sciences* **33**, 654–660.
URL: *https://doi.org/10.1016/j.proenv.2016.03.120*

Al-Areeq, A., Al-Zahrani, M. and Sharif, H. (2021), 'The performance of physically based and conceptual hydrologic models: A case study for makkah watershed, saudi arabia', *Water* **13**(8), 1098.
URL: *https://doi.org/10.3390/w13081098*

Bardossy, A., Bogardi, I. and Duckstein, L. (1990), 'Fuzzy regression in hydrology', *Water Resources Research* **26**(7), 1497–1508.
URL: *https://doi.org/10.1029/WR026i007p01497*

Beven, K. (1989), 'Changing ideas in hydrology — the case of physically-based models', *Journal of Hydrology* **105**(1–2), 157–172.
URL: *https://doi.org/10.1016/0022-1694(89)90101-7*

Du, H., Donat, M. G., Zong, S., Alexander, L. V., Manzanas, R., Kruger, A., Choi, G., Salinger, J., He, H. S., Li, M.-H., Fujibe, F., Nandintsetseg, B., Rehman, S., Abbas, F., Rusticucci, M., Srivastava, A., Zhai, P., Lippmann, T., Yabi, I. and Wu, Z. (2022), 'Extreme precipitation on consecutive days occurs more often in a warming climate', *Bulletin of the American Meteorological Society* **103**(4), E1130–E1145.
URL: *https://doi.org/10.1175/bams-d-21-0140.1*

Duncan, H. (2019), 'Baseflow separation – a practical approach', *Journal of Hydrology* **575**, 308–313.
URL: *https://doi.org/10.1016/j.jhydrol.2019.05.040*

Fadhil, M., Hidayat, Y. and Baskoro, D. (2021), 'Identifikasi perubahan penggunaan lahan dan karakteristik hidrologi das citarum hulu', *Jurnal Ilmu Pertanian Indonesia* **26**(2), 213–220.
URL: *https://doi.org/10.18343/jipi.26.2.213*

Furey, P. and Gupta, V. (2001), 'A physically based filter for separating base flow from streamflow time series', *Water Resources Research* **37**(11), 2709–2722.
URL: *https://doi.org/10.1029/2001WR000243*

Gelete, G., Nourani, V., Gokcekus, H. and Gichamo, T. (2023), 'Ensemble physically based semi-distributed models for the rainfall-runoff process modeling in the data-scarce katar catchment, ethiopia', *Journal of Hydroinformatics* **25**(2), 567–592.
URL: *https://doi.org/10.2166/hydro.2023.197*

Ginting, S. and Putuhena, W. (2017), 'Hujan rancangan berdasarkan analisis frekuensi regional dengan metode tl-moment', *JURNAL SUMBER DAYA AIR* **12**(1), 1–16.
URL: *https://doi.org/10.32679/jsda.v12i1.160*

Hou, W., Gao, J. and Wu, S. (2020), 'Quantitative analysis of the influencing factors and their interactions in runoff generation in a karst basin of southwestern china', *Water* **12**(10), 2898.
URL: *https://doi.org/10.3390/w12102898*

Jiang, S. and Kang, L. (2019), 'Flood frequency analysis for annual maximum streamflow using a non-stationary gev model', *E3S Web of Conferences* **79**, 03022.
URL: *https://doi.org/10.1051/e3sconf/20197903022*

Kobierska, F., Engeland, K. and Thorarinsdottir, T. (2018), 'Evaluation of design flood estimates – a case study for norway', *Hydrology Research* **49**(2), 450–465.
URL: *https://doi.org/10.2166/nh.2017.068*

Kuntoro, A., Cahyono, M. and Soentoro, E. (2018), 'Land cover and climate change impact on river discharge: Case study of upper citarum river basin', *Journal of Engineering and Technological Sciences* **50**(3), 364–381.
URL: *https://doi.org/10.5614/j.eng.technol.sci.2018.50.3.4*

Ladson, A., Brown, R., Neal, B. and Nathan, R. (2013), 'A standard approach to baseflow separation using the lyne and hollick filter', *Australian Journal of Water Resources* **17**(1).
URL: *https://doi.org/10.7158/W12-028.2013.17.1*

Mineo, C., Ridolfi, E., Neri, A. and Russo, F. (2019), 'Areal reduction factor: The effect of the return period', *AIP Conference Proceedings* p. 210004.
URL: *https://doi.org/10.1063/1.5114215*

Muhammad, A., Evenson, G., Stadnyk, T., Boluwade, A., Jha, S. and Coulibaly, P. (2019), 'Impact of model structure on the accuracy of hydrological modeling of a canadian prairie watershed', *Journal of Hydrology: Regional Studies* **21**, 40–56.
URL: *https://doi.org/10.1016/j.ejrh.2018.11.005*

Munier, S. and Decharme, B. (2022), 'River network and hydro-geomorphological parameters at 1/12° resolution for global hydrological and climate studies', *Earth System Science Data* **14**(5), 2239–2258.
URL: *https://doi.org/10.5194/essd-14-2239-2022*

Newman, A., Mizukami, N., Clark, M., Wood, A., Nijssen, B. and Nearing, G. (2017), 'Benchmarking of a physically based hydrologic model', *Journal of Hydrometeorology* **18**(8), 2215–2225.
URL: *https://doi.org/10.1175/JHM-D-16-0284.1*

Oktavia, S., Rustiati, N., Andiese, V., Amaliah, T., Labombang, M. and Mantika, O. (2022), 'Baseflow index on miu watershed based on a digital graphical method', *IOP Conference Series: Earth and Environmental Science* **1075**(1), 012052.
URL: *https://doi.org/10.1088/1755-1315/1075/1/012052*

Ostertagová, E. (2012), 'Modelling using polynomial regression', *Procedia Engineering* **48**, 500–506.
URL: *https://doi.org/10.1016/j.proeng.2012.09.545*

Paniconi, C. and Putti, M. (2015), 'Physically based modeling in catchment hydrology at 50: Survey and outlook', *Water Resources Research* **51**(9), 7090–7129.
URL: *https://doi.org/10.1002/2015WR017780*

Pavlovic, S., Perica, S., St Laurent, M. and Mejía, A. (2016), 'Intercomparison of selected fixed-area areal reduction factor methods', *Journal of Hydrology* **537**, 419–430.
URL: *https://doi.org/10.1016/j.jhydrol.2016.03.027*

Pietersen, J., Gericke, O., Smithers, J. and Woyessa, Y. (2015), 'Review of current methods for estimating areal reduction factors applied to south african design point rainfall and preliminary identification of new methods', *Journal of the South African Institution of Civil Engineering* **57**(1), 16–30.
URL: *https://doi.org/10.17159/2309-8775/2015/v57n1a2*

Poggio, L., de Sousa, L., Batjes, N., Heuvelink, G., Kempen, B., Ribeiro, E. and Rossiter, D. (2021), 'Soilgrids 2.0: producing soil information for the globe with quantified spatial uncertainty', *SOIL* **7**(1), 217–240.
URL: *https://doi.org/10.5194/soil-7-217-2021*

Potapov, P., Hansen, M., Pickens, A., Hernandez-Serna, A., Tyukavina, A., Turubanova, S., Zalles, V., Li, X., Khan, A., Stolle, F., Harris, N., Song, X.-P., Baggett, A., Kommareddy, I. and Kommareddy, A. (2022), 'The global 2000-2020 land cover and land use change

dataset derived from the landsat archive: First results', *Frontiers in Remote Sensing* **3**.
URL: *https://doi.org/10.3389/frsen.2022.856903*

Ratner, B. (2009), 'The correlation coefficient: Its values range between +1/−1, or do they?', *Journal of Targeting, Measurement and Analysis for Marketing* **17**(2), 139–142.
URL: *https://doi.org/10.1057/jt.2009.5*

Reis, D., Veilleux, A., Lamontagne, J., Stedinger, J. and Martins, E. (2020), 'Operational bayesian gls regression for regional hydrologic analyses', *Water Resources Research* **56**(8).
URL: *https://doi.org/10.1029/2019WR026940*

Rezaei, A., Ismail, Z., Niksokhan, M., Ramli, A., Sidek, L. and Dayarian, M. (2019), 'Investigating the effective factors influencing surface runoff generation in urban catchments – a review', *Desalination and Water Treatment* **164**, 276–292.
URL: *https://doi.org/10.5004/dwt.2019.24359*

Rusli, S., Weerts, A., Mustafa, S., Irawan, D., Taufiq, A. and Bense, V. (2023), 'Quantifying aquifer interaction using numerical groundwater flow model evaluated by environmental water tracer data: Application to the data-scarce area of the bandung groundwater basin, west java, indonesia', *Journal of Hydrology: Regional Studies* **50**, 101585.
URL: *https://doi.org/10.1016/j.ejrh.2023.101585*

Salim, A., Dharmawan, I. and Narendra, B. (2019), 'Pengaruh perubahan luas tutupan lahan hutan terhadap karakteristik hidrologi das citarum hulu', *Jurnal Ilmu Lingkungan* **17**(2), 333.
URL: *https://doi.org/10.14710/jil.17.2.333-340*

Samantaray, S. and Sahoo, A. (2020), 'Estimation of flood frequency using statistical method: Mahanadi river basin, india', *H2Open Journal* **3**(1), 189–207.
URL: *https://doi.org/10.2166/h2oj.2020.004*

Sebayang, I., Soekarno, I., Cahyono, M. and Kuntoro, A. (2022), 'Environmental flow assessment using low-flow index method in upper citarum river basin, west java, indonesia', *The Open Civil Engineering Journal* **16**(1).
URL: *https://doi.org/10.2174/18741495-v16-e221006-2022-2*

Shangguan, W., Dai, Y., Duan, Q., Liu, B. and Yuan, H. (2014), 'A global soil data set for earth system modeling', *Journal of Advances in Modeling Earth Systems* **6**(1), 249–263.
URL: *https://doi.org/10.1002/2013MS000293*

Smith, H. (1965), 'Handbook of applied hydrology. a compendium of water-resources technology. van te chow, ed. mcgraw-hill, new york, 1964. 1418 pp. illus.', *Science* **148**(3667), 219–219.
URL: *https://doi.org/10.1126/science.148.3667.219*

Sudradjat, A., Muhamad, B. and Nurohman, F. (2020), 'Contrasting climate induced variability of the upper citarum river baseflow and eventflow during early 20th century and recent decades', *E3S Web of Conferences* **148**, 03001.
URL: *https://doi.org/10.1051/e3sconf/202014803001*

Sun, J., Wang, X., Shahid, S. and Li, H. (2021), 'An optimized baseflow separation method for assessment of seasonal and spatial variability of baseflow and the driving factors', *Journal of Geographical Sciences* **31**(12), 1873–1894.
URL: *https://doi.org/10.1007/s11442-021-1927-8*

Sun, J., Ye, F., Nedjah, N., Zhang, M. and Xu, D. (2023), 'A practical yet accurate real-time statistical analysis library for hydrologic time-series big data', *Water* **15**(4), 708.
URL: *https://doi.org/10.3390/w15040708*

Taufik, M. and Annisa', S. (2022), 'Baseflow index analysis for bengawan solo river, indonesia', *Agromet* **36**(2), 70–78.
URL: *https://doi.org/10.29244/j.agromet.36.2.70-78*

Tunas, I. and Oka, G. (2020), 'Effect of data length to the consistency of design rainfall', *Journal of Physics: Conference Series* **1655**(1), 012119.
URL: *https://doi.org/10.1088/1742-6596/1655/1/012119*

van Kempen, G., van der Wiel, K. and Melsen, L. (2021), 'The impact of hydrological model structure on the simulation of extreme runoff events', *Natural Hazards and Earth System Sciences* **21**(3), 961–976.
URL: *https://doi.org/10.5194/nhess-21-961-2021*

Wang, F., Huang, G., Li, Y., Xu, J., Wang, G., Zhang, J., Duan, R. and Ren, J. (2021), 'A statistical hydrological model for yangtze river watershed based on stepwise cluster analysis', *Frontiers in Earth Science* **9**.
URL: *https://doi.org/10.3389/feart.2021.742331*

Wang, S., Peng, H., Hu, Q. and Jiang, M. (2022), 'Analysis of runoff generation driving factors based on hydrological model and interpretable machine learning method', *Journal of Hydrology: Regional Studies* **42**, 101139.
URL: *https://doi.org/10.1016/j.ejrh.2022.101139*

Wu, S.-J., Tung, Y.-K. and Yang, J.-C. (2009), 'Incorporating daily rainfall to derive at-site hourly depth-duration-frequency relationships', *Journal of Hydrologic Engineering* **14**(9), 992–1001.
URL: *https://doi.org/10.1061/(asce)he.1943-5584.0000065*

Yamazaki, D., Ikeshima, D., Tawatari, R., Yamaguchi, T., O'Loughlin, F., Neal, J., Sampson, C., Kanae, S. and Bates, P. (2017*a*), 'A high-accuracy map of global terrain elevations', *Geophysical Research Letters* **44**(11), 5844–5853.
URL: *https://doi.org/10.1002/2017GL072874*

Yamazaki, D., Ikeshima, D., Tawatari, R., Yamaguchi, T., O'Loughlin, F., Neal, J., Sampson, C., Kanae, S. and Bates, P. (2017*b*), 'A high-accuracy map of global terrain elevations', *Geophysical Research Letters* **44**(11), 5844–5853.
URL: *https://doi.org/10.1002/2017GL072874*

Yan, D., Li, C., Zhang, X., Wang, J., Feng, J., Dong, B., Fan, J., Wang, K., Zhang, C., Wang, H., Zhang, J. and Qin, T. (2022), 'A data set of global river networks and corresponding water resources zones divisions v2', *Scientific Data* **9**(1), 770.
URL: *https://doi.org/10.1038/s41597-022-01888-0*

Yoon, P., Kim, T.-W. and Yoo, C. (2013), 'Rainfall frequency analysis using a mixed gev distribution: a case study for annual maximum rainfalls in south korea', *Stochastic Environmental Research and Risk Assessment* **27**(5), 1143–1153.
URL: *https://doi.org/10.1007/s00477-012-0650-5*

Yulianto, F., Suwarsono, Nugroho, U., Nugroho, N., Sunarmodo, W. and Khomarudin, M. (2020), 'Spatial-temporal dynamics land use/land cover change and flood hazard mapping in the upstream citarum watershed, west java, indonesia', *Quaestiones Geographicae* **39**(1), 125–146.
URL: *https://doi.org/10.2478/quageo-2020-0010*

Zheng, Y., Li, S. and Ullah, K. (2020), 'Increased occurrence and intensity of consecutive rainfall events in the china's three gorges reservoir area under global warming', *Earth and Space Science* **7**(8).
URL: *https://doi.org/10.1029/2020ea001188*

Zhou, J., Zhao, C., Li, Q. and Yang, R. (2021), 'Research on baseflow separation based on single parameter digital filtering method', *IOP Conference Series: Earth and Environmental Science* **693**(1), 012079.
URL: *https://doi.org/10.1088/1755-1315/693/1/012079*