Journal of Hydrology: Regional Studies 54 (2024) 101899

Contents lists available at ScienceDirect

Journal of Hydrology: Regional Studies

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

Assessment of hydrological model performance in Morocco in relation to model structure and catchment characteristics

Oumar Jaffar [a,*], Abdessamad Hadri [a], El Mahdi El Khalki [a], Khaoula Ait Naceur [a], Mohamed Elmehdi Saidi [b], Yves Tramblay [c], Abdelghani Chehbouni [d]

[a] *International Water Research Institute (IWRI), Mohammed VI Polytechnic University (UM6P), Ben Guerir, Morocco*
[b] *L3G Laboratory, Department of Earth Sciences, Faculty of Sciences and Techniques, Cadi Ayyad University, Marrakech, Morocco*
[c] *HydroSciences Montpellier (University Montpellier, CNRS, IRD), Montpellier, France*
[d] *Center for Remote Sensing Applications (CRSA), Mohammed VI Polytechnic University (UM6P), Ben Guerir, Morocco*

A R T I C L E I N F O

Keywords:
Large-sample hydrology
Rainfall-runoff model
Hydrological modeling
Model complexity
Model intercomparison
Precipitation products
Rainfall forcing

A B S T R A C T

Study region: 30 catchments in Morocco.
Study focus: We assessed the KGE performance of eight monthly lumped hydrological models forced by ground-based rainfall observations. We then examined how the performance relates to model complexity and structure, applied exploratory correlation analysis to identify the catchment features (over 200 features were considered) most significantly related to model performance, and investigated how these models respond to three rainfall forcings (ERA5, CHIRPS, and PERSIANN-CDR).
New hydrological insights for the region: The findings indicate that no hydrological model outperformed (or underperformed) consistently across all the catchments and that model performance depends more on model structure and hydro-climatic characteristics, particularly those related to calibration and calibration-validation data difference, than on model complexity and non-hydro-climatic features. The linearity between rainfall and runoff was the primary feature influencing model performance. Additionally, besides the expected improvement of model performance when forced with richer rainfall and runoff calibration data in terms of wet and dry years, our results show that this holds true even if the calibration data is only relatively richer than the validation data and that dry periods are more beneficial to model performance than wet periods. Lastly, all the models responded similarly to the different rainfall inputs; each model performed better when using ERA5 than when using CHIRPS and underperformed when using PERSIANN-CDR. The metric that best explained this similarity was the Pearson correlation coefficient between the precipitation products and observed runoff.

1. Introduction

There is a growing interest in conducting large-sample hydrology studies (A Lane et al., 2019) – hydrological modeling studies involving a large sample of catchments. (Gupta et al., 2014) illustrated this tendency by providing a list of hydrological modeling studies that used 30 or more catchments. In recent years, several studies of this nature have been conducted in different parts of the

* Corresponding author.
E-mail address: oumar.jaffar@um6p.ma (O. Jaffar).

https://doi.org/10.1016/j.ejrh.2024.101899
Received 25 December 2023; Received in revised form 15 July 2024; Accepted 16 July 2024
Available online 28 July 2024

world, for example: (Kouakou et al., 2023) in Africa, (Z. Xu et al., 2022) in Asia, (Hapuarachchi et al., 2022) in Australia, (Pelletier and Andréassian, 2022) in Europe, (Fathi et al., 2023) in North America, and (Muñoz-Castro et al., 2023) in South America. This increase in popularity of large-sample modeling studies is unsurprising due to their usefulness in (i) detecting patterns about variables affecting model performance that could not have been possible to identify through studies based on a limited number of catchments (Gupta et al., 2014), (ii) deriving robust insights on hydrological processes (Addor et al., 2019), (iii) providing a valuable ground for models assessment and refining our understanding of models' functioning, strengths, and weaknesses (Knoben et al., 2020). (Andréassian et al., 2006) went so far as to state that this approach is a necessary condition to ensure progress in hydrology.

Models' assessment and model intercomparison studies across a large sample size of catchments can be tackled from different angles. This task can be undertaken, for example, by exploring the model performance dependency on model complexity and structure. (Mathevet et al., 2020) reported that the results of previous studies indicate that hydrological models can often yield comparable performances on a broad range of catchments, despite the differences in their structures and degrees of complexity. The finding concerning the model performance's lack of significant dependency regarding its complexity has been corroborated by many studies (Van Esse et al., 2013; Gan et al., 1997; Knoben et al., 2020; Perrin et al., 2001). In contrast, the model structure has been often reported to be a crucial element that explains model performance (Van Esse et al., 2013; Gan et al., 1997; Knoben et al., 2020; Perrin et al., 2001). From another models assessment angle, (A Lane et al., 2019; Massmann, 2020; McMillan et al., 2016; Pham et al., 2021; Poncelet et al., 2017; K. Zhang et al., 2023) examined the model performance relationship to catchment characteristics, and certain features emerged, frequently more than others, as being associated with model performance, such as: aridity index, baseflow index, and catchment area (Bai et al., 2015; Knoben et al., 2020; Massmann, 2020; Tramblay et al., 2023). Another common way of assessing hydrological models is by studying their responses when forced by different products of the model's inputs, for instance: (Bouizrou et al., 2023; Darbandsari and Coulibaly, 2020; Jayathilake and Smith, 2021; Oudin et al., 2005; J. Wang et al., 2023) employed different rainfall-runoff models forced using numerous precipitation and evapotranspiration inputs. The importance of this hydrological assessment has been recognized for quite some time as highlighted by (Stephens et al., 2022), particularly in areas with limited weather station density (S. Camici et al., 2018; Essou et al., 2016), since it is the practical approach to verify their adequacy for driving hydrological models.

In recent decades, a multitude of rainfall-runoff models have been developed, spanning from conceptual lumped models to physical-based distributed ones (Wan et al., 2021), and are used for a variety of purposes such as hydrological research (Jaiswal et al., 2020), water resources management (Dutta and Sarma, 2021), drought assessment (Mouelhi et al., 2006), and the assessment of impacts from climate change and anthropogenic activities (Bai et al., 2015), etc. However, (Andréassian et al., 2006) highlighted the impracticability of using physically-based models in model intercomparison studies, given that these studies require employing an automated process for dealing with the large number of catchments; a task that is relatively difficult to implement when using this type of models. Conversely, lumped models, monthly ones in particular, have landed themselves in the field of operational and comparative hydrology as a valuable tool for runoff prediction, especially in developing countries, such as Morocco, given their data-scarcity and data-sparsity situation (Adla et al., 2019; Bouizrou et al., 2023; Tramblay et al., 2016; Zamoum and Souag-Gamane, 2019). Additionally, given the drought period Morocco is currently immersed in along with being particularly exposed to the risks of climate change (Hadri et al., 2021), it is of great importance to assess various rainfall-runoff models across numerous Moroccan catchments.

In consideration of the foregoing, the primary objective of this article is to conduct a multi-faceted assessment of eight monthly lumped hydrological models on a set of 30 catchments in Morocco. To the best of our knowledge, the sample size of these catchments is the largest ever studied in Moroccan hydrological research using a similar number of hydrological models, and no research work has undertaken simultaneously both of the following tasks tackled in this study: (i) investigating a comparatively large set of features (over 200 features were investigated), and (ii) explicitly examining model performance's relation to both calibration data and the calibration-validation data difference. Regarding the specific research questions, they are detailed as follows: (i) Is there a best (or worst) performing model in all catchments? (ii) Is there a relationship between model performance and model complexity and structure? (iii) Which catchment features have the strongest influence on model performance? (iv) How does each model react to different rainfall products, and is there any metric that provides insight into which rainfall product will give better (or worse) hydrological model performance than the others?

To achieve the study's objectives, a model intercomparison is first carried out between eight conceptually diverse models with a varying number of free parameters ranging from 2 to 5 but with a similar spatial structure and number and timescale of inputs. The model conceptual structure as well as its complexity are then checked if they can explain the model performance differences – if such differences exist. Subsequently, a large set of catchment characteristics, including features that are calibration-period-related and calibration/validation-periods-related, is proposed to identify primary predictors of model performance. Finally, three rainfall products were used to force the hydrological models to examine how they respond to different precipitation inputs. Potential evapotranspiration input – the remaining necessary input required in most monthly lumped models (C. Y. Xu and Singh, 1998) – is not studied because previous studies (Jayathilake and Smith, 2021; Oudin et al., 2005) found that the lumped models they used displayed limited sensitivity to evapotranspiration forcing data, especially in water-limited catchments.

The paper is organized as follows: Section 2 describes the study area and data, Section 3 explains the methods, Section 4 presents and discusses the results, and Section 5 summarizes the findings and provides some practical implications.

2. Study area and data

2.1. Study area

This study was carried out in 30 catchments in Morocco (Fig. 1). The selected catchments were the result of the selection procedure of the analysis period (09/1983–08/2019) that was chosen to have the largest number of catchments not influenced by dams and with available rainfall and runoff data with, respectively, less than 20 % and 15 % as percentages of missing values.

The selected catchments and rain gauges resulted in a spatial configuration where the average area covered by each rain gauge station is about 680 Km^2 (with a maximum coverage of 2535 Km^2) and the average distance between each catchment centroid and its nearest rain gauge is 16 Km (with a maximum distance of 30 Km). This sample of catchments represents a varied range of physiographic and hydro-climatic characteristics as shown in Table 1:

The studied catchments have areas ranging from around 100 Km^2 to approximately 15000 Km^2. These catchments are all water-limited according to Budyko theory (P/PET < 1) (Budyko, 1958) and are dispersed in three climatic zones of the Köppen-Geiger classification system (Beck et al., 2018): Arid (class B), Temperate (class C), and Cold (class D), and three main land cover types according to Moderate Resolution Imaging Spectroradiometer (MODIS) Land Cover Type (MCD12Q1) (Sulla-Menashe and Friedl, 2022): Shrub, Grass, and Barren. It should be noted that the land cover class "Permanent Snow and Ice" is nonexistent in the study area, thus likely relatively limiting the snow's influence on catchment hydrology when compared to catchments with this cover class.

2.2. Rainfall data

2.2.1. Ground observations

A total of 34 rainfall stations (Fig. 1), were used to provide the monthly rainfall input for the hydrological models. A month was considered missing even with one missing daily value. Nonetheless, this resulted in data with limited missing values (the average missing months of all the stations is 5 %), and since all the investigated models necessitate complete data without any missing values, each gap in the monthly precipitation data was filled using simple linear regression with the station having the highest Pearson correlation coefficient (0.6 was the lower allowed threshold), which is a technique that belongs to the family of regression method commonly employed in the estimation of missing rainfall records (Teegavarapu et al., 2018; Teegavarapu and Chandramouli, 2005). The filled data were then used to calculate the weighted average rainfall records over the catchments by applying the Thiessen polygon method to the 34 rain gauge stations.

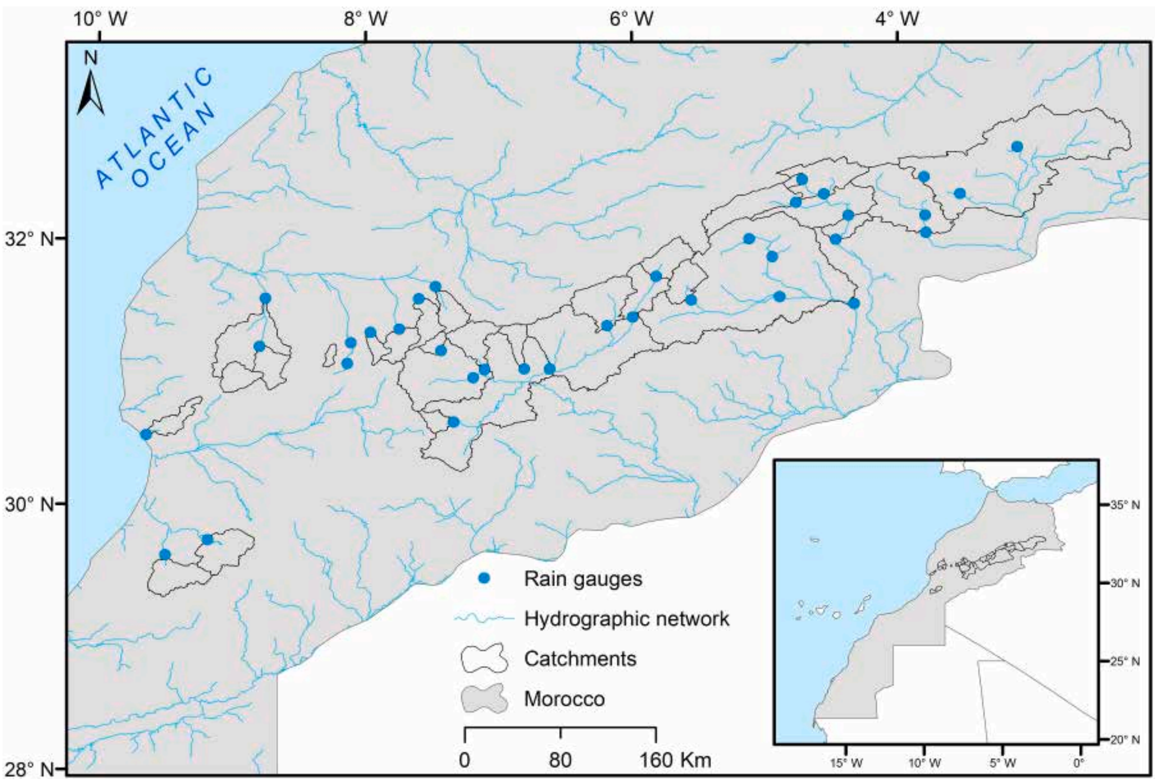

Fig. 1. Location of the 30 studied catchments and 34 rain gauge stations.

Table 1
Summary of the main characteristics of the 30 catchments included in this study.

Characteristic	Unit	10th percentile	Average	Median	90th percentile
Area	Km^2	222	2305	831	6667
Mean elevation	m	1250	1844	1863	2432
Mean slope	-	18 %	31 %	31 %	46 %
Mean annual precipitation (P)	mm	146	233	215	312
Mean annual runoff (Q)	mm	12	77	49	159
Mean annual potential evapotranspiration (PET)	mm	1042	1189	1208	1283
Mean annual temperature	°C	10.9	14.2	14.1	17.4
Aridity Index (P/PET)	-	0.11	0.20	0.18	0.26
Runoff Coefficient (Q/P)	-	0.07	0.30	0.24	0.52
Temporal average of Shrub cover percentages	-	16 %	42 %	32 %	84 %
Temporal average of Grass cover percentages	-	8 %	25 %	21 %	49 %
Temporal average of Barren cover percentages	-	2 %	31 %	26 %	71 %
Percentage of Arid (B) Climate	-	7 %	54 %	46 %	100 %
Percentage of Temperate (C) Climate	-	0 %	34 %	32 %	77 %
Percentage of Cold (D) Climate	-	0 %	12 %	4 %	42 %

The data sources of the mentioned characteristics will be presented in Section 2.2, Section 2.3, Section 2.4, and Section 2.5.

2.2.2. *Other rainfall products*

Three other precipitation products were used as forcing data for the hydrological models only to examine the models' responses to diverse rainfall input data, and were not used in any other analysis. These products belong to three datasets denoted here as ERA5, CHIRPS, and PERSIANN-CDR, and are briefly described in Table 2. The selection of these products was based mainly on their temporal coverage since they cover the period of analysis (09/1983–08/2019), and also because they were previously evaluated in Morocco (e.g., (Ait Dhmane et al., 2023; El Khalki et al., 2023; Najmi et al., 2023; Ouatiki et al., 2023; Rachdane et al., 2022; Salih et al., 2022)). The records of these products were converted into monthly time-steps and the data pixels within each catchment were averaged to derive the mean areal rainfall input for the hydrological models.

2.3. *Potential evapotranspiration data*

Monthly potential evapotranspiration data is the second and last required input for the studied hydrological models. This input was computed with the Hargreaves-Samani formula (Hargreaves and Samani, 1985) requiring latitude as well as daily maximum, minimum, and average air temperature that were retrieved from ERA5 product, and spatially averaged at each catchment scale. The use of ERA5 temperature data was motivated by its suitable temporal coverage and previous utilization for hydrological modeling both in Morocco and throughout Africa (Tarek et al., 2020, 2021; Tramblay et al., 2023). The choice of the Hargreaves-Samani formula on the other hand is supported by its comparatively good performance in computing potential evapotranspiration in arid and semi-arid regions in Morocco (Er-Raki et al., 2010) and worldwide (Pimentel et al., 2023), and by the fact that over two-thirds (69 %) of the study area falls under the category of arid regions according to the Köppen-Geiger classification. The daily data output by this formula was then converted into monthly time-steps.

2.4. *Observed streamflow data*

Concerning observed streamflow data, 30 gauging stations that represent the outlets of the 30 studied catchments were used for the calibration and validation of the hydrological models. Similar to precipitation data, the collected discharge data featured a mix of daily and monthly records, and the daily records were aggregated to generate monthly ones, while considering a month as missing if one or more of its daily values are missing. On average, the percentage of months with missing discharge data is 7 %.

2.5. *Catchment features*

To study how models' performance is related to catchments' characteristics, 227 features were selected in this study, belonging to 8 categories:

- Morphometric, Topographic, and Hydrographic Network Characteristics (17 features).
- Catchment Coordinates and its Relative Location to Rainfall Stations [used to provide its precipitation input] (7 features).
- Hydro-Climatic Features (59 features).
- Hydro-Climatic Features in Calibration (59 features).
- Hydro-Climatic Features Disparity: Calibration vs. Validation (59 features). Each of the characteristics of this category is calculated as the feature's value during calibration minus its value in the validation period.
- Climate Classes (9 features).
- Land Cover Classes (7 features, including the temporal change in the land cover classes).
- Soil Textures and Properties (10 features).

Table 2
Description of the three employed precipitation products.

Notation	ERA5	CHIRPS	PERSIANN-CDR
Temporal coverage	1979-present	1981-present	1983-present
Spatial resolution	0.25°	0.05°	0.25°
Description	Precipitation is obtained using data assimilation by combining the physical-based integrated forecasting system of ECMWF with observed data	Precipitation results from the incorporation of NASA and NOAA gridded satellite-based precipitation estimates with in-situ precipitation data	Precipitation is developed by the University of California and is derived using an artificial neural networks algorithm that utilizes gridded satellite (GridSat-B1) infrared data.
Reference	(Hersbach et al., 2020)	(Funk et al., 2015)	(Ashouri et al., 2015)

The exhaustive list of the used features can be found in Appendix A.

The datasets used for catchment features determination are given in Table 3:

Additionally, hydro-climatic features included the diversity of temperature data, characterized by the presence of cold and hot years, as well as the richness of precipitation and runoff data, characterized by wet and dry years. Z-score was employed to classify the years, with −1 and 1 being used as threshold values distinguishing between cold/wet, normal, and hot/dry categories, as it is commonly the case for many drought indices and studies (Elair et al., 2023; Jain et al., 2015; Mahmoudi et al., 2019; Salih et al., 2023).

3. Methods

The methodology proposed for tackling the aforementioned research questions is presented in the following flowchart(Fig. 2)

3.1. Hydrological models

The eight monthly lumped hydrological models selected for comparison are listed in Table 4:

It should be noted that the references from which the VUB model was derived, present more than one way for runoff computation depending on multiple combinations of mathematical equations used for the calculation of evapotranspiration, direct and indirect runoff components. The VUB equations used in this study are the ones that yielded the best validation results in the studied catchments and are provided in Appendix B.

All the employed hydrological models incorporate a soil moisture storage. GR2M and GR5M have an additional routing storage, while abcd, DWBM, and Wapaba integrate a groundwater storage. The remaining models do not feature any additional reservoir. The 2-parameter models have one runoff component, direct runoff which in this case is equivalent to total runoff, and the remaining models have an additional component: indirect runoff.

Additionally, the selected models require monthly inputs, which are more available than daily data, especially in developing countries (Adla et al., 2019). These models were chosen with the primary aim of studying the impact of model complexity – defined here as the number of model parameters – and model structure while isolating the effects of spatial structure and the inputs' number, types, and timescales. The ease of the model's utilization was also considered, with a preference for lumped models known for their simplicity and suitability for operational hydrology tasks (Mouelhi et al., 2006; Perrin et al., 2001; Tramblay et al., 2016).

Among the selected hydrological models in this study, GR2M model is by far the most used in Morocco, particularly for climate change impact assessment studies (Driouech et al., 2010; Hrour et al., 2023; El Khalki et al., 2021; Saouabe et al., 2022).

3.2. Models' calibration and evaluation

The eight hydrological models were calibrated using the Differential Evolution optimization algorithm (Storn and Price, 1997) with the Kling-Gupta efficiency (KGE) used as an objective function (Gupta et al., 2009). In its formulation, KGE is based on three equally weighted components: Pearson correlation (r) for matching the timing and shape of the observed hydrograph, bias (β) for reproducing the observed runoff volume, and variability ratio (α) for capturing the observed flows' spread:

$$KGE = 1 - \sqrt{(1-r) + (1-\beta) + (1-\alpha)} \tag{1}$$

with: $\beta = \mu_s(Q)/\mu_o(Q)$ and $\alpha = \sigma_s(Q)/\sigma_o(Q)$; $\mu_o(Q)$ and $\sigma_o(Q)$ represent, respectively, the mean and standard deviation of the observed runoff, while $\mu_s(Q)$ and $\sigma_s(Q)$ represent, respectively, the mean and standard deviation of the simulated runoff.

The period of analysis is composed of 36 hydrological years of monthly data, with each hydrological year spanning from 1st September to 31st August. One year served as a warm-up period and the remaining years were divided into two independent periods: calibration period and validation period. Comparable to the odd/even split-sample testing used for example by (Arsenault et al., 2016;

Table 3
Datasets used for catchment features determination.

Category of features	Dataset used	Reference
Morphometric, Topographic, and Hydrographic Network Characteristics	Shuttle Radar Topography Mission (SRTM) Global 1 arc second.	(Farr et al., 2007)
Hydro-Climatic Features	ERA5 was used to retrieve temperature data, which was then used for the estimation of potential evapotranspiration.	(Hersbach et al., 2020)
	Ground observational data were used to calculate features related to precipitation and runoff.	-
Land Cover Classes	Yearly data of MODIS Land Cover Type (MCD12Q1) from the period 2001–2019.	(Sulla-Menashe and Friedl, 2022)
Climate Classes	Köppen-Geiger classification.	(Beck et al., 2018)
Soil Textures and Properties	iSDAsoil dataset was used to retrieve soil textures at 0–20 cm and 20–50 cm depths.	(Hengl et al., 2021)
	Field capacity and Saturated hydraulic conductivity were estimated using average values based on soil textures.	(Th van Genuchten et al., 1991) (Gong et al., 2012)

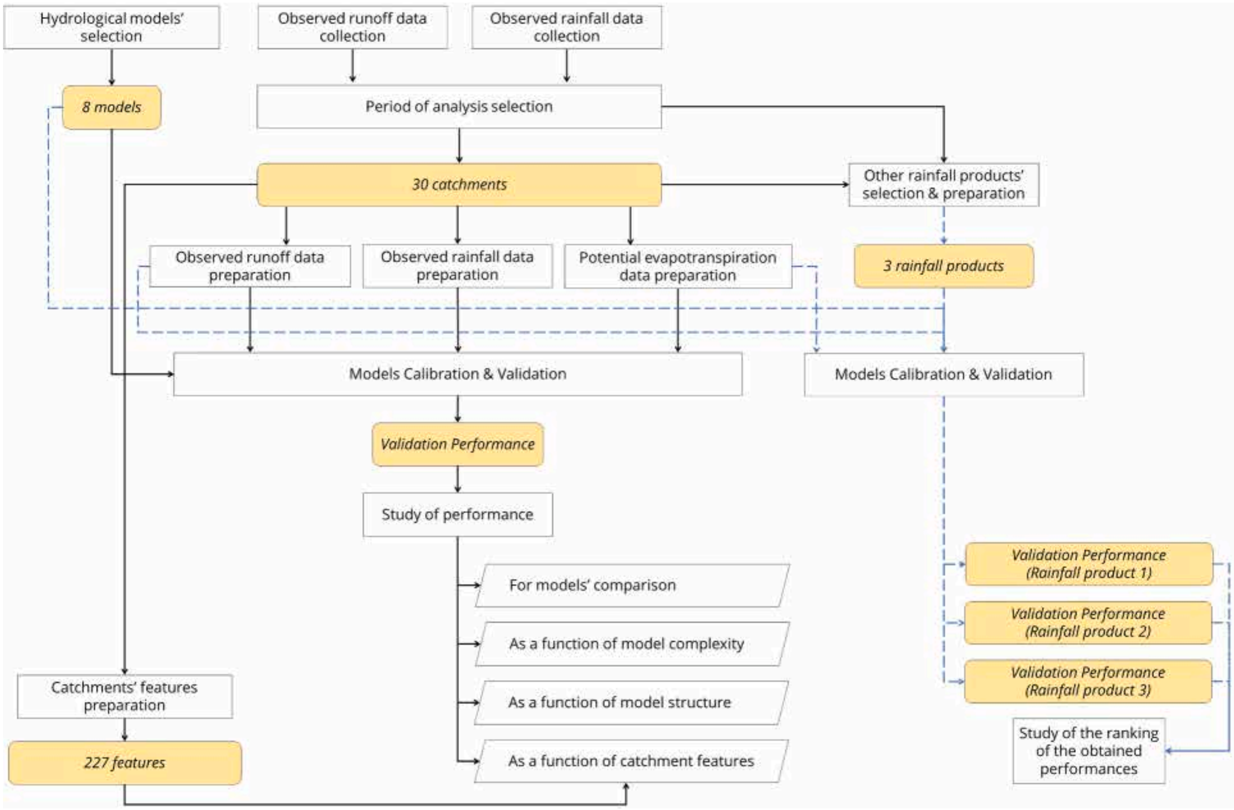

Fig. 2. Flowchart of the adopted methodology.

Essou et al., 2016; Gowda et al., 2012), a similar approach was employed in which each period (calibration and validation) is formed by uniting sets consisting of two consecutive years, with each set being separated from the next by two consecutive years as shown in Fig. 3.

The hydrological models' performances were assessed in the validation period because the validation mode is generally the most used mode in operational contexts (Perrin et al., 2001) in addition to the fact that the performance obtained in the calibration period is potentially subject to the problem of over-fitting that can result from over-parameterization (Seibert et al., 2019). This performance was assessed by employing KGE since it provides a more comprehensive evaluation of the overall hydrograph (Ahmed et al., 2023) and it is currently considered to be the most used and recommended criterion for assessing the performance of hydrological models (Stefania Camici et al., 2020). Additionally, the KGE bias component β was only used to investigate whether the models manifest a bias towards overestimating or underestimating runoff, without utilizing it for comparing the models.

3.3. Models' performance predictors identification

The focus of this task is on identifying catchment features that influence the validation performance. This identification was accomplished using exploratory correlation analysis between the models' performance and the catchments' characteristics, following the same approach used by (Kim and Lee, 2014; Knoben et al., 2020; Massmann, 2020; Pham et al., 2021). Spearman correlation was adopted for this analysis. The influence of a given feature on the performance metric (KGE) was considered significant only if the p-value of the corresponding Spearman correlation coefficient was lower than 0.05 (a threshold that is commonly used in hydrological studies (Poncelet et al., 2017)) for at least half of the employed hydrological models. This latter condition was considered to minimize the possibility of maintaining, and therefore studying, weak relationships between models' performance and the identified features.

4. Results and discussion

The results presented in Sections 4.1 through 4.4 were obtained using observed rainfall as forcing data for the hydrological models.

4.1. Models' performance assessment and comparison

Fig. 4 illustrates the spatial distribution of KGE in the validation period for all the hydrological models. This figure shows a large spread of performance across the studied catchments, for all the investigated models, ranging from below 0 to more than 0.75, with a

Table 4
Characteristics of the 8 studied hydrological models."X" and "-" indicate respectively the existence and absence of the corresponding characteristic.

Model			GR2M	XM	WM	VUB	abcd	DWBM	GR5M	Wapaba
Reference			(Mouelhi et al., 2006)	(Xiong and Guo, 1999)	(G. Q. Wang et al., 2014)	(G. L. Vandewiele and Ni-Lar-Win, 2009; G. Vandewiele and Xu, 1992)	(Thomas Jr. Harold, 1981)	(L. Zhang et al., 2008)	(Mouelhi et al., 2006)	(Q. J. Wang et al., 2011)
Number of parameters			2	2	3	3	4	4	5	5
Storages	**Number of storages**		2	1	1	1	2	2	2	2
	Type of storages	**Soil moisture storage**	X	X	X	X	X	X	X	X
		Routing storage	X	-	-	-	-	-	X	-
		Groundwater storage	-	-	-	-	X	X	-	X
Runoff components	**Number of runoff components**		1	1	2	2	2	2	2	2
	Type of runoff components	**Direct runoff**	X	X	X	X	X	X	X	X
		Indirect runoff	-	-	X	X	X	X	X	X
Non-linearity of total runoff generation with regard to variables	**Precipitation**		X	X	-	X	X	X	X	X
	Potential evapotranspiration		X	X	-	X	-	X	X	X
	Soil moisture storage level		X	X	-	X	X	X	X	X
Number of parameters included in direct runoff calculation			2	2	1	1	3	2	4	3
Number of parameters included in total runoff calculation			2	2	2	2	4	3	5	4

Fig. 3. Calibration/Validation setup.

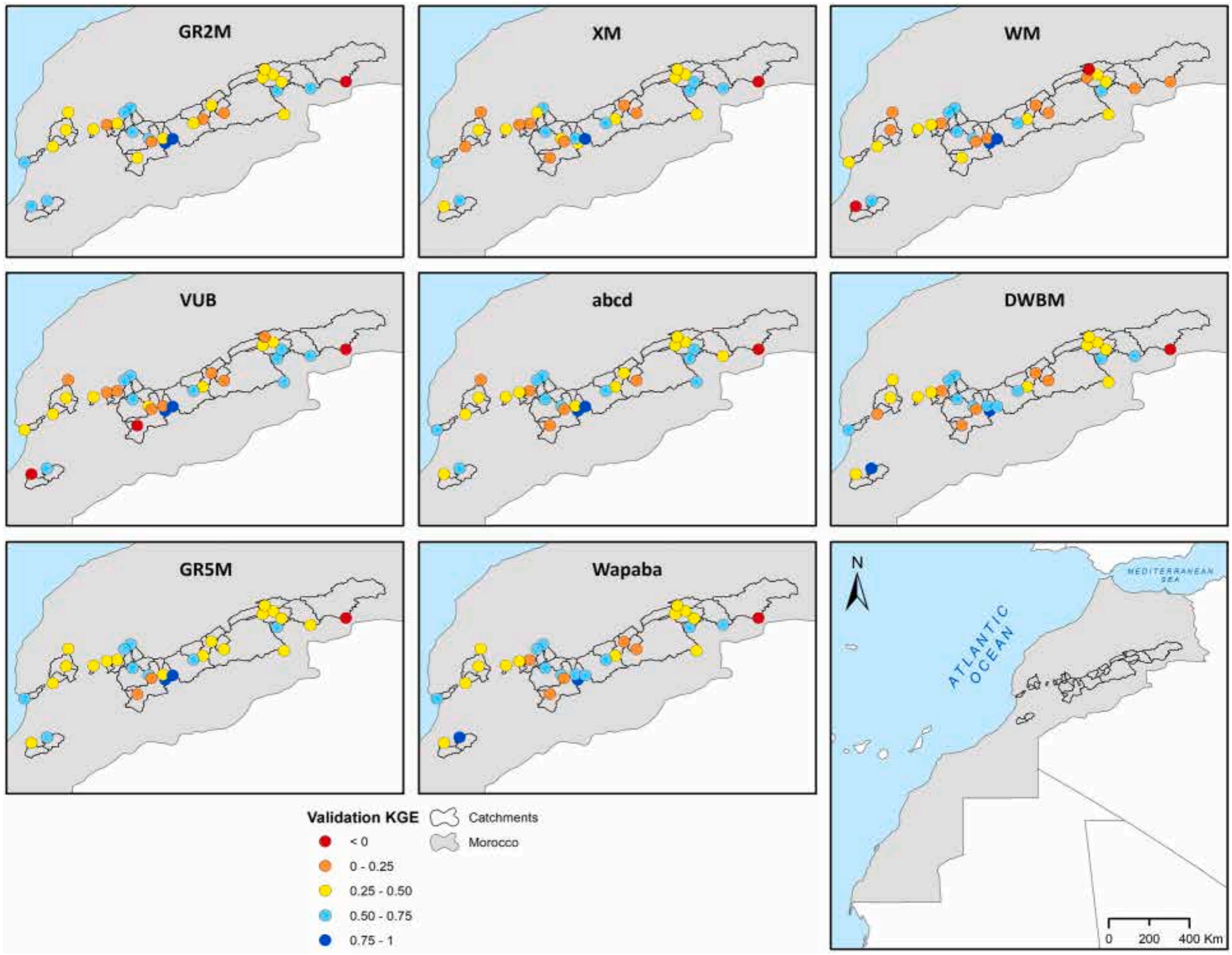

Fig. 4. Spatial distribution of the eight models' validation performances (KGE) at the studied catchments' outlets.

predominance of the yellow (0.25 < KGE < 0.50) and white-blue (0.50 < KGE < 0.75) categories. Additionally, and consistently across all models, three is the highest number of catchments with a negative KGE and two is the maximum number of catchments with KGE > 0.75. While there is no noticeable spatial trend of performance aligned with any specific geographic orientation for all eight models, the models nonetheless exhibit different levels of performance as further shown in Fig. 5.

The calibration performance results can be found in Appendix C.

Fig. 5 presents several statistics detailing the models' validation performance within the study area. In terms of median KGE performance, WM exhibited the lowest value (0.36) while the other seven models showed comparable performances with medians ranging from 0.42 (VUB) to 0.48 (XM). As for the average KGE performance of the models, XM, WM, and VUB demonstrated similar results with a mean of 0.36 which is lower than the remaining models' average performance of 0.43, with Wapaba having the maximum KGE average. However, as displayed in Table 5, it should be noted that no hydrological model was the best in all catchments

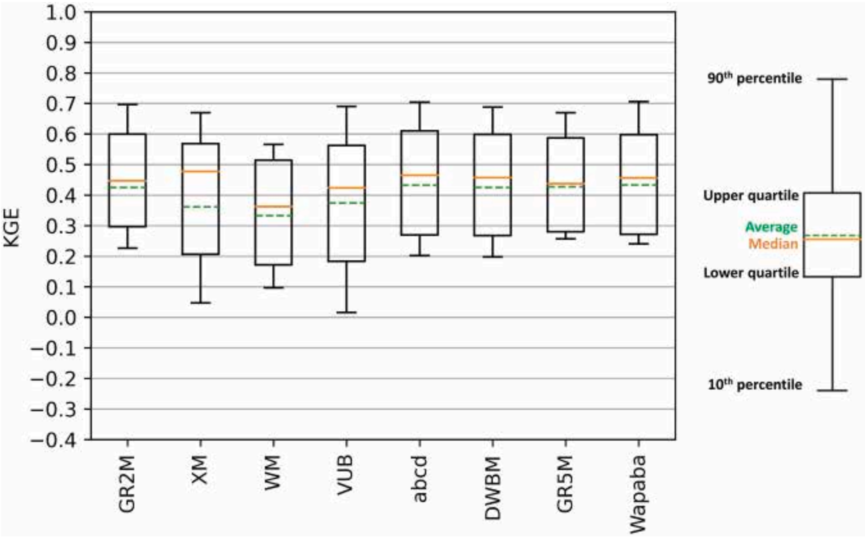

Fig. 5. Boxplots of KGE validation performance in the different studied catchments for all the hydrological models. As illustrated on the right side of the figure, and in all the figures containing boxplots throughout the paper, the orange line represents the median, the dotted green line represents the average, the box edges represent the 25th and 75th percentiles, and the whiskers represent the 10th and 90th percentiles.

(all percentages <100 %). Moreover, each model was the best in at least three catchments (10 %) with GR5M being the best-performing model. These results are in line with (Bai et al., 2015) in their comparison of monthly water balance models. More generally, the observation of the nonexistence of a best-performing model across all catchments was also identified by (Knoben et al., 2020; Orth et al., 2015; Perrin et al., 2001) in their comparative assessments of daily hydrological models. (De la Fuente et al., 2023) even recommended abandoning the concept of a "best" model and suggested focusing instead on learning from the strengths of each model representation. Furthermore, this nonexistence of a best-performing model might have partially accounted for the rationale behind the development of blended models (e.g., (Chlumsky et al., 2021)) that calculate each modeled hydrological process as a weighted average of different mathematical formulations, not just a single one as it is the case for most conceptual rainfall-runoff models, thus providing a greater adjustment capability to suit a larger variability of catchments. Another finding is that no hydrological model was the worst in all catchments (all percentages <100 %) and each model, except for Wapaba, was the worst in at least 1 catchment (3 %) with WM being the worst-performing model.

Another criterion that can be used for models' comparison is the percentage (or number) of catchments for which a hydrological model achieved a KGE score exceeding the 0.5 threshold. While any threshold can serve for the comparison, we opted for 0.5 due to its prior utilization as a KGE threshold between "good" and "bad" hydrological performance in a large-sample hydrological study (Stefania Camici et al., 2020). Based on this criterion, XM, DWBM, and Wapaba are the top-ranked (Table 5).

From the above results (Table 5) and analysis, it is apparent that while using only one performance metric KGE, the hydrological models can be compared using different criteria (median performance, average performance, percentage of catchments with best model performance, percentage of catchments with worst model performance, percentage of catchments with model performance > threshold), each of which may result in a distinct model ranking as shown in Table 5.

For this reason, we opted for the introduction of a new aggregated comparative measure, which we have named the Multi-Criteria Comparative Performance Score (MCCPS), and believe that it provides a more encompassing comparison between the models. For each hydrological model, MCCPS is calculated as the mean of the normalized values (normalized using the min-max normalization) of the five used comparison criteria in Table 5, namely: the median KGE, the average KGE, the percentage of catchments for which a model demonstrated the best KGE performance, the percentage of catchments for which a model demonstrated the worst KGE performance, and the percentage of catchments for which a model scored a KGE above 0.5. Therefore, MCCPS ranges potentially from 0 to 1: a score of 1 indicates that the model has achieved the best value for all the five used criteria, whereas a score of 0 suggests that the model has obtained the worst value for these criteria. The MCCPS of the investigated models is displayed in Fig. 5.

Fig. 6 illustrates that based on MCCPS, Wapaba is the best-ranked model, followed by abcd and GR5M which yielded comparable results. DWBM and GR2M constitute the second-best group with homogenous scores. WM however is by far the worst-performing model with a MCCPS of zero, meaning that this model had the lowest median performance, the lowest average performance, the lowest percentage of catchments with best model performance, the lowest percentage of catchments with worst model performance, and the lowest percentage of catchments with model performance > 0.5. WM had nonetheless the best model performance in three catchments.

In relation to the question of whether the employed hydrological models overestimate or underestimate the runoff, Fig. 7 shows that for all the models, the validation KGE bias component β at 80 % of the catchments is between 0.7 and 1.4 with median and average values ranging from 0.9 to 1.1, indicating that the models on average and for 50 % of the catchments reproduce reasonably the total

Table 5

Values of five model comparison criteria during the validation period for the eight models, along with their rankings based on each criterion. These criteria are: median KGE – Criterion-1, average KGE – Criterion-2, percentage of catchments for which the model demonstrated the best KGE performance – Criterion-3, percentage of catchments for which the model demonstrated the worst KGE performance – Criterion-4, and percentage of catchments for which the model scored a KGE greater than 0.5 – Criterion-5.

	Criterion-1	Rank according to Criterion-1	Criterion-2	Rank according to Criterion-2	Criterion-3	Rank according to Criterion-3	Criterion-4	Rank according to Criterion-4	Criterion-5	Rank according to Criterion-5
GR2M	0.447	5	0.425	4	23 %	2	13 %	5	37 %	6
XM	0.478	1	0.361	7	17 %	4	33 %	7	43 %	1
WM	0.363	8	0.333	8	10 %	6	40 %	8	30 %	8
VUB	0.424	7	0.374	6	10 %	6	13 %	5	40 %	4
abcd	0.465	2	0.432	2	20 %	3	3 %	2	40 %	4
DWBM	0.457	3	0.425	5	10 %	6	7 %	4	43 %	1
GR5M	0.438	6	0.427	3	30 %	1	3 %	2	37 %	6
Wapaba	0.456	4	0.433	1	17 %	4	0 %	1	43 %	1

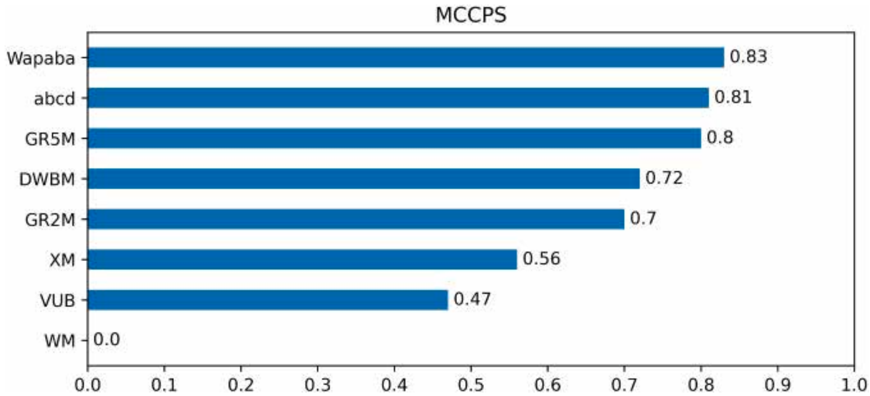

Fig. 6. MCCPS of the eight hydrological models. The models are sorted in descending order from top to bottom.

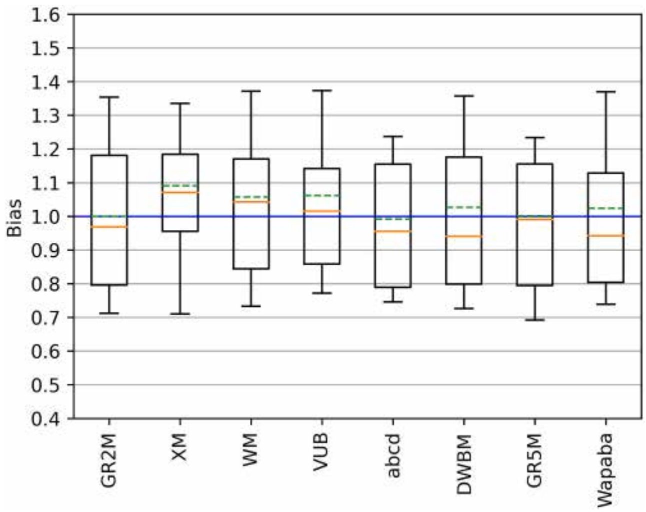

Fig. 7. Boxplots of the KGE bias component β (in validation period) in the different studied catchments for all the hydrological models.

runoff volume with the exception of XM that comparatively demonstrates a tendency to overestimation.

The MCCPS is the measure of model performance that will be used for studying the models' performance with respect to model complexity and structure. It is important to highlight that the observations as well as the [non]significance of the relationships reported below remain valid in the case of choosing the average KGE, instead of MCCPS, as the measure of overall model performance.

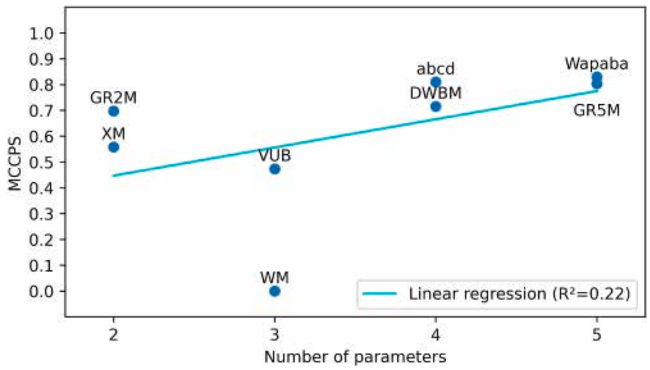

Fig. 8. Model performance in validation as a function of model complexity.

4.2. Models' performance with regard to model complexity

Given the difference in performance between the eight hydrological models, it was essential to try to identify how much of this disparity can be explained by the models' complexity. Adopting the point of view that relates model complexity solely to the number of free parameters in calibration (e.g., (Bai et al., 2015; Gan et al., 1997; Mathevet et al., 2020; Perrin et al., 2001; Schoups et al., 2008), Fig. 8 presents the model performance as a function of complexity:

Even though the performance of the 4 and 5-parameter models is relatively high, Fig. 8 suggests that the relation between model performance and model complexity is not significant (p-value = 0.144 > 0.05; threshold value that has been selected to distinguish between significant and non-significant relations), which is clearly illustrated by the relatively high performance of the 2-parameter models GR2M and XM compared to the 3-parameter models WM and VUB. Our results agree with (Bai et al., 2015; Gan et al., 1997; Knoben et al., 2020; Mathevet et al., 2020; Perrin et al., 2001) who concluded that, for both the employed daily and monthly models, an increase in model complexity does not necessarily guarantee an increase in model performance, and simple models can yield comparable results to those of more complex models. (Massmann, 2020) argued that this can be explained either (i) by the over-parameterization issue generally encountered in more complex models, which in turn could result from the unsuitability of these models' structures as suggested by (Perrin et al., 2001), or (ii) by the fact that the increase in model complexity did not tackle the underlying reason behind the poor performance of a simpler model, which hints at the possible increase in model performance if the shortages in the processes representation are corrected by the additional parameters as shown in the work of (Massmann, 2020), particularly in the arid cluster, in which the increase in model complexity translated into an increase in model performance. This was likely the case in their study because the employed models were related to one another in that models with more parameters contained, in a sense, the models with fewer parameters as a special case, thus potentially addressing the causes of the suboptimal performance observed in simpler models, which is not the case for the eight models in our study that are structurally distinct with parameters that generally have different purposes in each model.

This finding suggests the existence of stronger explanatory elements to the models' performance difference other than the number of model parameters and the model structure is the main and foremost point of consideration.

4.3. Models' performance with regard to model structure

To study the models' performance with respect to model structure, we investigated the existence of possible relations between model performance, represented by the MCCPS, and elements from Table 4 and found that:

- The 2-storage models (GR2M, abcd, DWBM, GR5M, and Wapaba) perform better than the 1-storage models (XM, WM, and VUB). This observation is inconsistent with earlier findings by (Bai et al., 2015; Knoben et al., 2020; Massmann, 2020) in their model intercomparison studies, in which no apparent trend was observed between the number of storages and model performance.
- The number of runoff components does not seem to have any influence on model performance, which corroborates the results obtained by (Bai et al., 2015).
- The number of parameters included in total runoff calculation is not significantly related to model performance.
- The relationship between model performance and the "number of parameters included in direct runoff calculation" is positive and significant (p-value = 0.025 < 0.05) (Fig. 9), at least to a certain threshold number.

The inconsistency in the relationship's significance between the number of storages and the model performance could have stemmed from the small sample size of just two categories: 1-storage models and 2-storage models even though eight models were used, and/or from the fact that the number of storages alone is insufficient to explain the variation in model performance.

As for the number of parameters utilized in computing direct runoff, to our knowledge, there is no study that has investigated the relationship between this variable and model performance, and therefore further research is needed to confirm or refute the

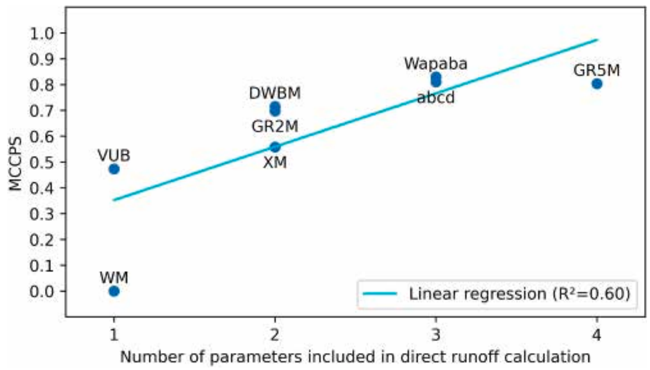

Fig. 9. Model performance in validation as a function of the number of parameters included in direct runoff calculation.

significance of this relationship, as our findings highlight its significance.

As mentioned earlier, WM model has a MCCPS of zero, indicating that this model has the worst rank with regard to median performance, average performance, percentage of catchments with best model performance, percentage of catchments with worst model performance, and percentage of catchments with model performance > 0.5. We argue that this comparatively weak performance is a result of WM using only one parameter for direct runoff computation and being the only model that computes runoff as a linear function of precipitation and soil moisture storage level. The latter argument is probably the main cause behind its poor performance since, as highlighted by (Rajurkar et al., 2004), simple linear relationships in hydrological models fail to capture the non-linear dynamics that are intrinsic in rainfall-runoff transformation, especially in ephemeral streams with a low yield (Ye et al., 1997) as those present in our study area. This argument is supported by the comparative study of conceptual model structures done by (Van Esse et al., 2013) in France, who found that linear runoff functions presented a lesser performance than that of non-linear ones.

In addition, it is noteworthy that the 2-parameter model GR2M displayed a relatively good performance. We hypothesize that one of the factors contributing to this level of performance is the comparatively detailed description of the hydrological processes which is reflected, for example, in the integration of two storages and in the number of values taken by each state variable "soil moisture storage level" and "routing storage level" in a given time-step —three values, a distinction that sets GR2M apart from the other models except GR5M, as they all use only one value per time-step.

The results of these analyses support the perspective that model performance is affected more by model structure and the way runoff generation is described than by increasing the total number of model parameters (Van Esse et al., 2013; Gan et al., 1997; Perrin et al., 2001), defined here as model complexity.

4.4. Models' performance with regard to catchment features

As mentioned before, the relation between model performance in the validation period and catchment features was investigated using the Spearman correlation. Among the eight used categories of catchment features (see Section 2.5), only those related to hydro-climatic characteristics had significant relations (p-value < 0.05) for at least half of the models (Fig. 10) (these features are presented in Fig. 11). More specifically and for all the hydrological models, the calibration/validation-related hydro-climatic features constitute from 50 % to 80 % of the features with significant predictive power for hydrological performance (KGE). Pearson correlation between monthly rainfall and observed streamflow is the only hydro-climatic feature, independent of the calibration and validation period, that showed a significant correlation with KGE.

It is worth noting that certain features were identified as significant based on the Spearman correlation coefficient's value but were nonetheless excluded from the analysis. This decision was based on a visual inspection of the data using scatterplots, revealing no discernible relationship between these features and KGE performance, similar to the example illustrated in Fig. 4 of Anscombe's quartet (Anscombe, 1973).

The observed absence of significant correlations between model performance and other catchment features such as topographic, landscape, and soil properties aligns with the findings of (Knoben et al., 2020) who found, using correlation analysis, that the attributes related to topography, soil, and vegetation are among the weakest predictors of model performance. (Massmann, 2020) found that this is the case only for certain models. These results can be explained by (i) the non-use of the suitable performance metric as (McMillan et al., 2016) found that only some performance metrics were sensitive to catchment characteristics, (ii) the fact that the observed hydrograph is a complex result of multiple features, as argued by (Knoben et al., 2020), which challenges the isolation of the effect of individual features, (iii) the non-use of the appropriate catchment features that possibly have significant relationships to KGE, and/or by (iv) the possibility that the used catchments have limited variability with regard to these features. An example of the latter argument is the aridity index, often identified as a significant factor influencing model performance (Bai et al., 2015; De la Fuente et al., 2023;

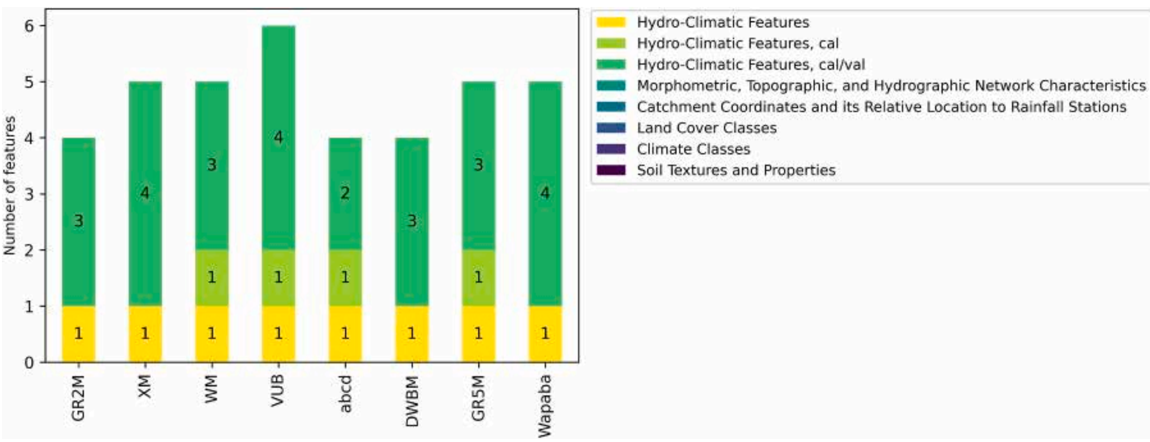

Fig. 10. Types and numbers of features with significant relationships (p-value < 0.05), for at least half of the models, with model performance (KGE) in validation.

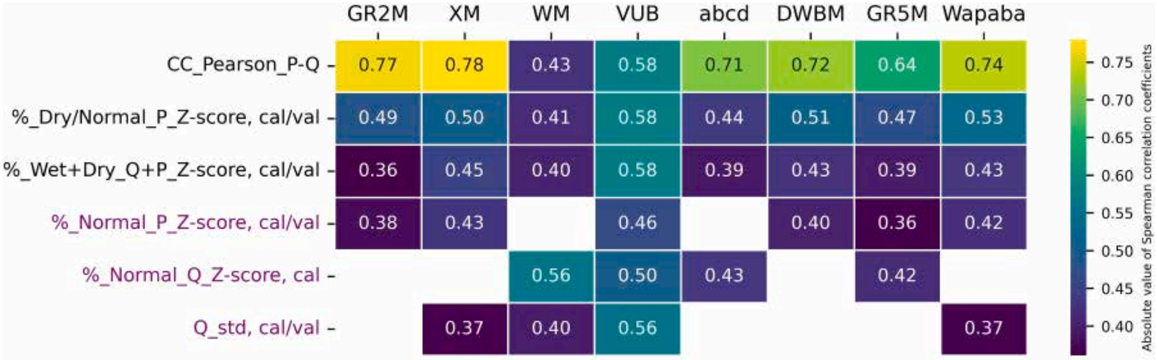

Fig. 11. Absolute value of Spearman correlation between model performance (KGE) in validation and catchment features. Only correlations with p-value < 0.05 for at least half of the models are shown. The characteristics in violet present a negative correlation, while those in black exhibit a positive one. These features are described in Appendix A.

Massmann, 2020; Poncelet et al., 2017), which is not the case in our study area, likely because all catchments have an aridity index below 0.5.

Regarding the observed significance of the period-dependent hydro-climatic features in relation to model performance, this result supports the findings of (Bisselink et al., 2016; Mathevet et al., 2020) for the calibration-related features, and (Coron et al., 2012; Dakhlaoui et al., 2017; Merz et al., 2011; Motavita et al., 2019) for calibration/validation-related features, highlighting the model performance dependency on hydro-climatic characteristics in calibration and validation periods.

Fig. 11 presents the details of Fig. 10. These results demonstrate that the models respond similarly to the variations of some features but differently to others, presumably due to the differences in the models' structures. Nevertheless, the findings highlight that:

– The linearity between rainfall and runoff is the most significant factor influencing model performance and proves beneficial for all the hydrological models; the stronger this correlation, the easier the catchment is modeled. This corroborates the findings of (Bisselink et al., 2016) who concluded, using 7 precipitation products as inputs for one daily model applied to four catchments, that model performance tends to improve with increasing linearity between monthly rainfall and observed runoff, on the condition that rainfall variability remains within acceptable limits.
– The calibration data richness in terms of its content of wet and dry years relative to precipitation and runoff is of great significance to model performance; a conclusion that was previously stated by (Gan et al., 1997) who attributed good model performance to data containing both low and high flows which provide sufficient information for the models during calibration. In line with this discussion, (Jakeman and Hornberger, 1993) attributed the high likelihood of reliable parameter estimation to the richness of information contained in the used data in terms of hydrologic variability. Our results also show that, interestingly, this remains valid even if the calibration data is only relatively richer than the validation data and that dry conditions, specifically those related to precipitation, seem to have a positive greater impact on the performance compared to wet conditions. This latter observation agrees with (Li et al., 2012; Vaze et al., 2010) findings even in Australia's context, and hints at the more useful information contained in dry periods, compared to wet periods, for model calibration.
– We found that if the annual runoff variability – represented by the annual runoff standard deviation – is larger during calibration than in the validation period, the hydrological model tends to perform poorly. We hypothesize that this is the case for certain models due to the possibility that they may capture random fluctuations in the calibration data that do not necessarily relate to the underlying hydrological processes. Relevant to this observation, (Poncelet et al., 2017) revealed that, independently of the calibration and validation periods, the annual variability of runoff flashiness has a negative impact on model performance.

Another factor that influences catchment hydrology, and would potentially affect model performance, is the process of snowmelt (Boudhar et al., 2009; Hanich et al., 2022; López-Moreno et al., 2023; Singh and Muñoz-Arriola, 2021). For this reason, the catchments were selected with no permanent snow cover class to limit the snow influence since the hydrological models were assessed without a snowmelt module. The choice of not using a snow component aligns with the main objective of this work which is to evaluate the performance of the hydrological models as they are in terms of complexity and structure. Nonetheless, since some studied catchments belong to the Atlas Mountains, we investigated the effect of snowmelt on model performance by comparing the average KGE of the snowmelt-affected catchments and the non-snowmelt-affected ones. The result of this comparison and the classification method of the snowmelt-affected catchments are provided in the Supplementary Material. Fig. S1 in Supplementary Material shows that model performance, on average, is worse in snowmelt-affected catchments than in the non-snowmelt-affected catchments. This finding highlights the importance, particularly for hydrological prediction, of the numerous research efforts carried out for the estimation of snow-related parameters in the Atlas Mountains (Baba et al., 2021; Bouamri et al., 2021; Bousbaa et al., 2022; Marchane et al., 2015; Rhoujjati et al., 2023) and points to the possible added value of including a snow component in hydrological models when the primary research objective is the study of individual mountainous catchments or the improvement of their discharge simulations.

4.5. Models' performance with regard to different precipitation forcing data

The models' performance dependency on precipitation-related features further emphasizes the need for evaluating the models' responses to various rainfall products. So, besides the observed rainfall input, the eight hydrological models were forced by monthly precipitation from ERA5, CHIRPS, and PERSIANN-CDR, and the obtained results in the validation period are displayed in Fig. 12.

From a first visual inspection of this figure, it appears that, for all the models, the overall performance degrades gradually when forced by observed rainfall to ERA5, then to CHIRPS, and finally to PERSIANN-CDR. To verify this observation, we used the same metric MCCPS previously employed to evaluate and compare the models with each other, yet this time, the objective is to compare, for each model, the responses of the different precipitation products used.

Fig. 13 confirms the finding that all the used hydrological models react similarly to the various precipitation products in question and demonstrates that for the eight hydrological models used, the overall performance obtained when using ERA5 precipitation product as forcing data is better than that of CHIRPS and that PERSIANN-CDR yields the poorest hydrological performance. This figure also shows that the observed rainfall remains, despite the sparsity of the rainfall network, the most valuable source of input for the hydrological models in the study area.

It is important to note that, here again, this ranking remains valid in the case of using the average KGE as the performance metric instead of MCCPS.

In an attempt to explain the consistency in the ranking of each model's responses corresponding to the three precipitation products, we employed seven continuous statistical indices:

- Six metrics, commonly used in rainfall products evaluation studies (Benkirane et al., 2023; Stefania Camici et al., 2020; Ouatiki et al., 2017; Rachidi et al., 2023; Saouabe et al., 2020), relating precipitation products to observed rainfall at the catchment scale: Kling-Gupta efficiency score (KGE_P-Pobs), Pearson correlation coefficient (CC_P-Pobs), a bias-related metric (Bias*, see Table 6), root mean square error (RMSE_P-Pobs), mean absolute error (MAE_P-Pobs), and Nash-Sutcliffe efficiency (Nash and Sutcliffe, 1970) (NSE_P-Pobs).
- One metric relating precipitation products to observed runoff: Pearson correlation coefficient (CC_P-Qobs).

It is important to note that the evaluation of the precipitation products is beyond the scope of this paper. For information purposes, the boxplots of these seven indices, for ERA5, CHIRPS, and PERSIANN-CDR, across the 30 studied catchments can be found in Appendix D.

For all the models, we investigated for each catchment the compatibility between (i) the ranking of the precipitation products with regard to each metric, and (ii) the ranking of models' responses to these products when used as forcing data. Specifically for each model, a catchment receives a compatibility score of 1 for a given metric if the ranking of the precipitation products according to that metric matches the ranking of the hydrological performances associated with these products. Otherwise, the compatibility score is 0. For each metric and model, we subsequently calculate the cumulative compatibility score by summing the number of catchments that

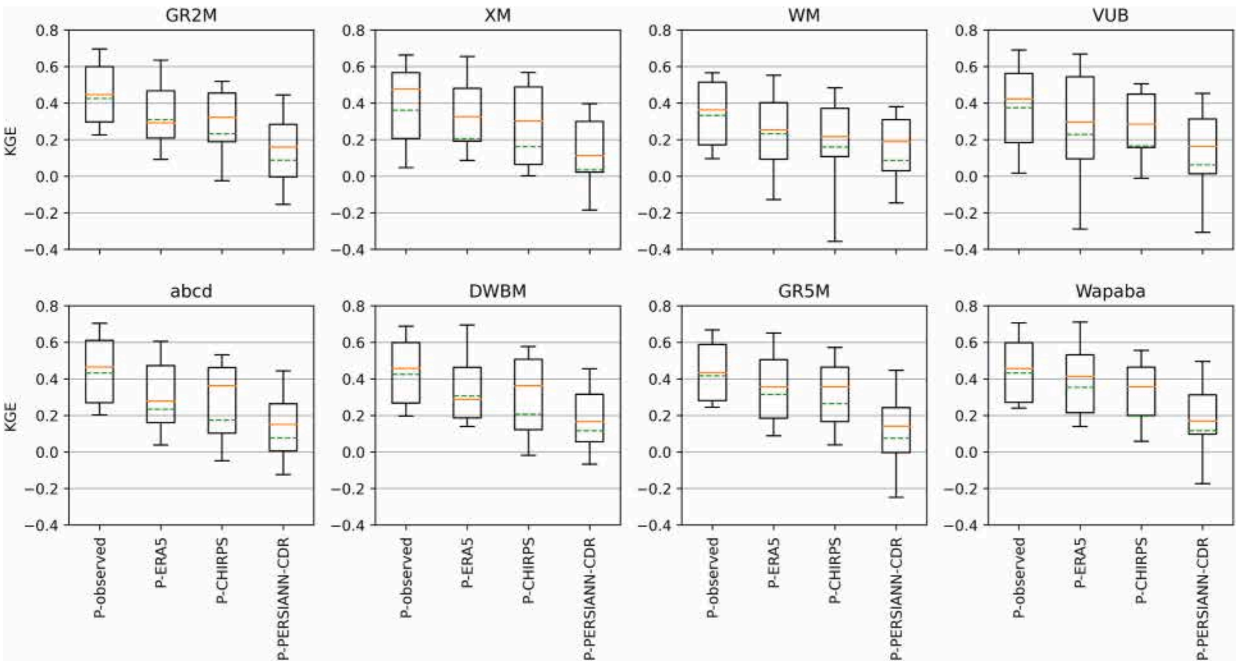

Fig. 12. Models' performance during validation as forced by observed rainfall, ERA5, CHIRPS, and PERSIANN-CDR precipitation products.

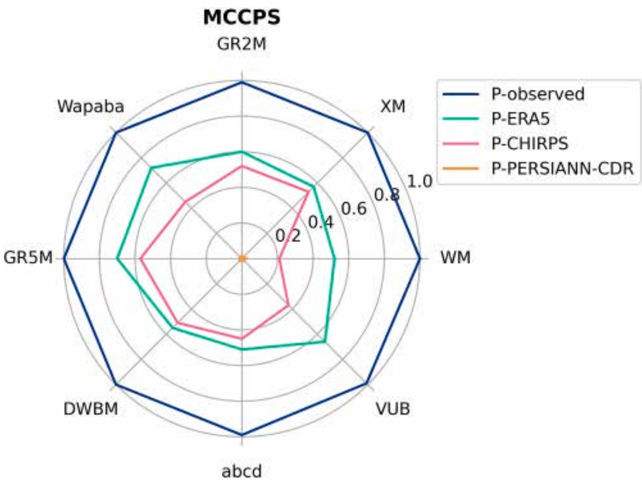

Fig. 13. Radar chart of models' performances in validation when forced using observed rainfall, ERA5, CHIRPS, and PERSIANN-CDR precipitation products.

Table 6
Bias-related metric used for precipitation products assessment.

Notation	Formula
Bias*	$Bias^{*} = 1 - \|1 - Bias\|$ (2)
Bias	$Bias = \mu_S(P)/\mu_o(P)$ (3)

$\boldsymbol{\mu_o}(\boldsymbol{P})$ represent the mean of the observed rainfall while and $\boldsymbol{\mu_s}(\boldsymbol{P})$ represent the mean of the simulated rainfall.

demonstrate compatibility with respect to that specific metric. Therefore, for each model, the cumulative compatibility score of a given metric ranges potentially from 0 to 30, and the metric with the highest cumulative score is considered the most explanatory metric for the hydrological performance of the precipitation products.

Fig. 14 shows that Pearson correlation between precipitation products and observed runoff (CC_P-Qobs) is by far the most explanatory index for the uniformity in the models' behavior when forced by the various precipitation products since it has the highest cumulative compatibility score for 7 out of 8 models, meaning that the superior hydrological performance of ERA5 (and of CHIRPS compared to PERSIANN-CDR) for all the models is due to its relatively high linear correlation with the observed streamflow of the studied catchments. This finding is supported by the evidence we found in the previous analysis in which we found that this feature is the most important predictor for model performance, which reinforces the similar results found by (Bisselink et al., 2016). However,

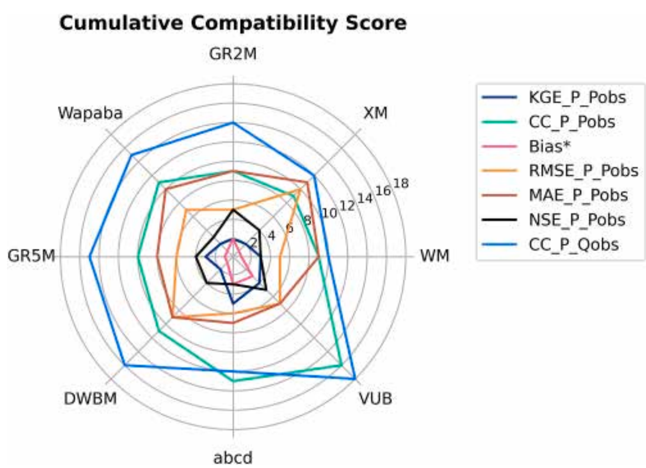

Fig. 14. Radar chart of the cumulative compatibility score of the seven used statistical metrics for all the hydrological models.

this result does not align with the findings of (Stefania Camici et al., 2020) in which they concluded, using a single daily model that was forced by inputs from three precipitation products across more than 1300 catchments in Europe, that RMSE_P-Pobs and Bias are the most reliable indices to select the best-performing precipitation product for hydrological modeling. We hypothesize that this was the case either (i) because, for each used product, RMSE_P-Pobs and Bias were monotonically correlated to CC_P-Qobs in the context of their study area, or (ii) because the structure of MISDc (Brocca et al., 2011) – the daily hydrological model used in their study – differs greatly from that of the eight monthly models we employed.

5. Conclusions

The aim of this large-sample hydrology study was to assess the performance of eight monthly lumped hydrological models (GR2M, XM, WM, VUB, abcd, DWBM, GR5M, and Wapaba) in 30 Moroccan catchments. As part of this assessment, we examined how model performance, expressed as KGE, relates to model complexity, model structure, and catchment features (more than 200 features were used), and we investigated how the employed models react to three rainfall products (ERA5, CHIRPS, and PERSIANN-CDR) when used as forcing input data. The main findings can be summarized as follows, with each accompanied by its practical implication:

- Across all catchments, no hydrological model exhibited consistently superior or inferior performance. Additionally, using a new comparative measure (MCCPS) that offers a more encompassing comparison between the models, Wapaba was found to be the best-ranked model in terms of overall performance, followed by abcd and GR5M that yielded comparable results, then by DWBM, GR2M, XM, VUB, and WM that scored a null MCCPS. *Implication*: Therefore, when possible, it is important to consider more than one hydrological model in runoff prediction.
- Model performance is affected more by model structure and how runoff generation is described than by model complexity. In particular, the number of storages as well as the number of parameters included in direct runoff calculation seemed to affect positively the model performance. Additionally, the relatively good performance of the 2-parameter model GR2M and bad performance of the 3-parameter model WM were attributed respectively to (i) the detailed description of the hydrological processes and (ii) the simplistic linear representation of the runoff generation process with regard to precipitation. *Implication*: Therefore, in model development, it is crucial to place greater emphasis on the mathematical equations used to describe runoff generation – with the avoidance of linear relationships between runoff and precipitation – rather than solely relying on using more free parameters.
- Contrary to non-hydro-climatic factors such as the physiographic characteristics, hydro-climatic features especially the calibration-related and the calibration/validation-related features are the only type of characteristics that had a statistically significant link to model performance. *Implication*: Therefore, in large-sample hydrology studies, it is worthwhile to test various calibration/validation scenarios for each of the studied catchments, and then apply the most suitable scenario for each catchment. This proposed approach is in opposition to the conventional one employed in this study which uses a single scenario for all the catchments.
- Model performance is significantly and positively related to the richness of rainfall and runoff calibration data regarding its content of wet and dry years, both in absolute terms and in calibration relative to validation, with the particularity of the dry conditions being more impactful than the wet ones. *Implication*: It is then important to find the most relevant index, to model performance, to characterize this richness.
- In response to varied rainfall forcing data, all the models exhibited a consistent pattern: superior performance with ERA5 compared to CHIRPS, and lower performance with PERSIANN-CDR. *Implication*: Within the study area and in situations where observed rainfall data is unavailable and there is limited capacity to analyze hydrological model responses to different rainfall products, the use of ERA5 precipitation product is recommended.
- The primary factor that significantly impacts model performance is the linear relationship between rainfall and runoff; a stronger linear correlation between these two variables results in enhanced model performance. *Implication*: Thus, in hydrological applications, it is advantageous for the performance of hydrological models to base the choice of (i) a method for filling missing rainfall values, (ii) a technique for spatially interpolating rainfall data, or (iii) a precipitation product for forcing hydrological models, on the highest linear correlation with the observed runoff.

CRediT authorship contribution statement

Oumar Jaffar: Writing – review & editing, Writing – original draft, Methodology, Formal analysis, Data curation, Conceptualization. **Abdessamad Hadri:** Writing – review & editing, Supervision, Methodology, Conceptualization. **El Mahdi El Khalki:** Writing – review & editing, Supervision, Methodology, Data curation, Conceptualization. **Khaoula Ait Naceur:** Writing – review & editing, Data curation. **Mohamed Elmehdi Saidi:** Writing – review & editing. **Yves Tramblay:** Writing – review & editing. **Abdelghani Chehbouni:** Writing – review & editing, Supervision, Methodology.

Declaration of Competing Interest

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

Data Availability

Data will be made available on request.

Appendix A. List of catchment Features

Table A1
List of the used catchment features in the correlation analysis.

Category of feature	Feature	Abbreviation	Unit
Morphometric, Topographic, and Hydrographic Network Characteristics	Area	A	Km^2
	Perimeter	Pr	Km
	Gravelius Compactness Index	Gc	-
	Maximum Catchment Elevation	Z_max	m
	Mean Catchment Elevation	Z_mean	m
	Median Catchment Elevation	Z_median	m
	Minimum Catchment Elevation	Z_min	m
	Catchment Elevation Range	Z_range	m
	Mean Catchment Slope	S_mean	-
	Median Catchment Slope	S_median	-
	Longest Flow Path	Stream_L	Km
	Stream Slope	Stream_S	-
	Maximum Stream Order	SO_max	-
	Total Length of Streams	L_total	Km
	Drainage Density	DD	Km/Km^2
	Hypsometric Integral	HI	-
	Time of Concentration	Tc	h
Catchment Coordinates and its Relative Location to Rainfall Stations	Catchment Latitude	Latitude_C	°
	Catchment Longitude	Longitude_C	°
	Outlet Latitude	Latitude_O	°
	Outlet Longitude	Longitude_O	°
	Number of Rainfall Stations	NB_stations	-
	Rain Gauge Density	Stations_density	Km^2
	Distance to the Nearest Station	D_nearest_station	Km
Hydro-Climatic Features	Mean Annual Precipitation	P	mm
	Annual Precipitation Standard deviation	P_std	mm
	Annual Precipitation Coefficient of variation	P_cv	-
	Percentage of Wet years with regard to Precipitation	%_Wet_P_Z-score	-
	Percentage of Dry years with regard to Precipitation	%_Dry_P_Z-score	-
	Percentage of Normal years with regard to Precipitation	%_Normal_P_Z-score	-
	Dry Precipitation-Wet Precipitation difference	%_Dry_P-Wet_P_Z-score	-
	Wet Precipitation-Normal Precipitation ratio	%_Wet/Normal_P_Z-score	-
	Dry Precipitation-Normal Precipitation ratio	%_Dry/Normal_P_Z-score	-
	Winter Rainfall	P_winter	mm
	Spring Rainfall	P_spring	mm
	Summer Rainfall	P_summer	mm
	Fall Rainfall	P_fall	mm
	Percentage of missing monthly runoff	%_Missing_Q	-
	Mean Annual runoff	Q	mm
	Annual runoff Standard deviation	Q_std	mm
	Annual runoff Coefficient of variation	Q_cv	mm
	Percentage of Wet years with regard to Runoff	%_Wet_Q_Z-score	-
	Percentage of Dry years with regard to Runoff	%_Dry_Q_Z-score	-
	Percentage of Normal years with regard to Runoff	%_Normal_Q_Z-score	-
	Dry Runoff-Wet Runoff difference	%_Dry_Q-Wet_Q_Z-score	-
	Wet Runoff-Normal Runoff ratio	%_Wet/Normal_Q_Z-score	-
	Dry Runoff-Normal Runoff ratio	%_Dry/Normal_Q_Z-score	-
	Winter Runoff	Q_winter	mm
	Spring Runoff	Q_spring	mm
	Summer Runoff	Q_summer	mm
	Fall Runoff	Q_fall	mm
	Combined Wet and Dry percentages for Precipitation and Runoff	%_Wet+Dry_Q+P_Z-score	-
	Annual Potential Evapotranspiration	PET	mm
	Winter Potential Evapotranspiration	PET_winter	mm
	Spring Potential Evapotranspiration	PET_spring	mm

(continued on next page)

Table A1 (*continued*)

Category of feature	Feature	Abbreviation	Unit
	Summer Potential Evapotranspiration	PET_summer	mm
	Fall Potential Evapotranspiration	PET_fall	mm
	Mean annual Temperature	T	°C
	Annual Temperature Standard deviation	T_std	°C
	Annual Temperature Coefficient of variation	T_cv	-
	Percentage of Cold years with regard to Temperature	%_Cold_T_Z-score	-
	Percentage of Hot years with regard to Temperature	%_Hot_T_Z-score	-
	Percentage of Normal years with regard to Temperature	%_Normal_T_Z-score	-
	Hot Temperature -Cold Temperature difference	%_Hot_T-Cold_T_Z-score	-
	Cold Temperature -Normal Temperature ratio	%_Cold/Normal_T_Z-score	-
	Hot Temperature -Normal Temperature ratio	%_Hot/Normal_T_Z-score	-
	Winter Temperature	T_winter	°C
	Spring Temperature	T_spring	°C
	Summer Temperature	T_summer	°C
	Fall Temperature	T_fall	°C
	Combined Richness for Precipitation and Temperature	%_Wet+Dry_P+%_Cold+Hot_T_Z-score	-
	Aridity Index	P/PET	-
	Winter Aridity Index	P/PET_winter	-
	Spring Aridity Index	P/PET_spring	-
	Summer Aridity Index	P/PET_summer	-
	Fall Aridity Index	P/PET_fall	-
	Runoff Coefficient	Q/P	-
	Winter Runoff Coefficient	Q/P_winter	-
	Spring Runoff Coefficient	Q/P_spring	-
	Summer Runoff Coefficient	Q/P_summer	-
	Fall Runoff Coefficient	Q/P_fall	-
	Pearson correlation between monthly Precipitation and Runoff	CC_Pearson_P-Q	-
	Ratio of the standard deviation of annual runoff and precipitation	Q_std/P_std	-
Hydro-Climatic Features in Calibration	The Hydro-Climatic Features during calibration. They are labeled using the same notations as the Hydro-Climatic Features, with an added ", cal" suffix.		
Hydro-Climatic Features Disparity: Calibration vs. Validation	The Hydro-Climatic Features' values during calibration minus their values in the validation period. They are labeled using the same notations as the Hydro-Climatic Features, with an added ", cal/val" suffix.		
Land Cover Classes	Temporal average of Shrub cover percentages	Shrub_%	-
	Temporal average of Grass cover percentages	Grass_%	-
	Temporal average of Barren cover percentages	Barren_%	-
	Temporal standard deviation of Shrub cover percentage	Shrub_std	-
	Temporal standard deviation of Grass cover percentage	Grass_std	-
	Temporal standard deviation of Barren cover percentage	Barren_std	-
	Mean absolute year-to-year fluctuations of Shrub cover percentage	Shrub_MAy2y	-
	Mean absolute year-to-year fluctuations of Grass cover percentage	Grass_MAy2y	-
	Mean absolute year-to-year fluctuations of Barren cover percentage	Barren_MAy2y	-
Climate Classes	Percentage of Arid Desert Climate	BW_Climate_%	-
	Percentage of Arid Steppe Climate	BS_Climate_%	-
	Percentage of Temperate Climate with Dry Summer	Cs_Climate_%	-
	Percentage of Cold Climate with Dry Summer	Ds_Climate_%	-
	Percentage of Arid Climate	B_Climate_%	-
	Percentage of Temperate Climate	C_Climate_%	-
	Percentage of Cold Climate	D_Climate_%	-
Soil Textures and Properties	Percentage of Clay Loam soil texture at 0–20 cm depth	ClayLoam_0–20 cm_%	-
	Percentage of Sandy Clay Loam soil texture at 0–20 cm depth	SandyClayLoam_0–20cm_%	-
	Percentage of Sandy Loam soil texture at 0–20 cm depth	SandyLoam_0–20cm_%	-
	Percentage of Clay Loam soil texture at 20–50 cm depth	ClayLoam_20–50 cm_%	-
	Percentage of Sandy Clay Loam soil texture at 20–50 cm depth	SandyClayLoam_20–50cm_%	-
	Percentage of Sandy Loam soil texture at 20–50 cm depth	SandyLoam_20–50cm_%	-

(*continued on next page*)

Table A1 (*continued*)

Category of feature	Feature	Abbreviation	Unit
	Field Capacity at 0–20 cm	FieldCapacity_0–20 cm	-
	Field Capacity at 20–50 cm	FieldCapacity_20–50 cm	-
	Saturated Hydraulic Conductivity at 0–20 cm	Ks_0–20 cm	mm/h
	Saturated Hydraulic Conductivity at 20–50 cm	Ks_20–50 cm	mm/h

Appendix B. Runoff and evapotranspiration equations used in VUB model

$$S_{t-1}^{+} = \max(S_{t-1}, \quad 0) \tag{B.1}$$

$$ET_t = PET_t \times \left(1 - \quad a_3^{\frac{\left(P_t + S_{t-1}^{+}\right)}{PET_t}}\right) \tag{B.2}$$

$$Q_t = \quad a_1 \quad \times \quad S_{t-1}^{+} + a_2 \quad \times \quad \sqrt{S_{t-1}^{+}} \times \left(P_t - PET_t \times \left(1 - \quad e^{-\frac{P_t}{PET_t}}\right)\right) \tag{B.3}$$

$$S_t = S_{t-1} + P_t - ET_t - Q_t \tag{B.4}$$

S_t: Soil moisture storage level, S_{t-1}: Soil moisture storage level of the previous month,

ET_t: Evapotranspiration, PET_t: Potential evapotranspiration, P_t: Precipitation

a_1, a_2, and a_3 are the model's parameters and represent respectively the slow runoff coefficient, fast runoff coefficient, and evapotranspiration coefficient.

Appendix C. Calibration KGE performance in the different studied catchments for all the hydrological models

.

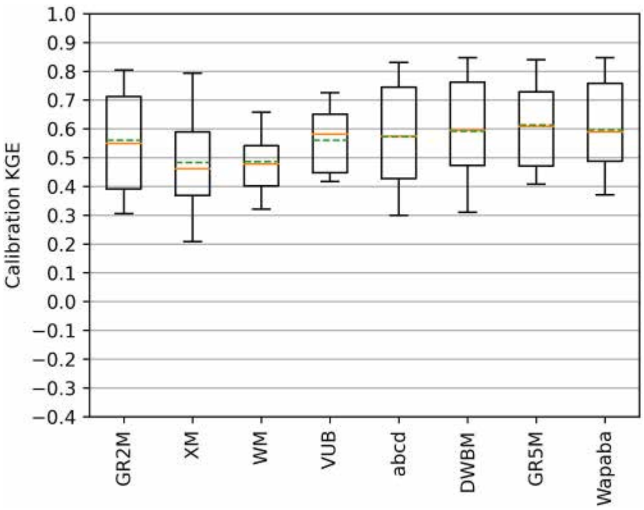

Fig. C.1. Boxplots of KGE calibration performance in the different studied catchments for all the hydrological models.

Appendix D. Results of precipitation products assessment at the catchments' scale

.

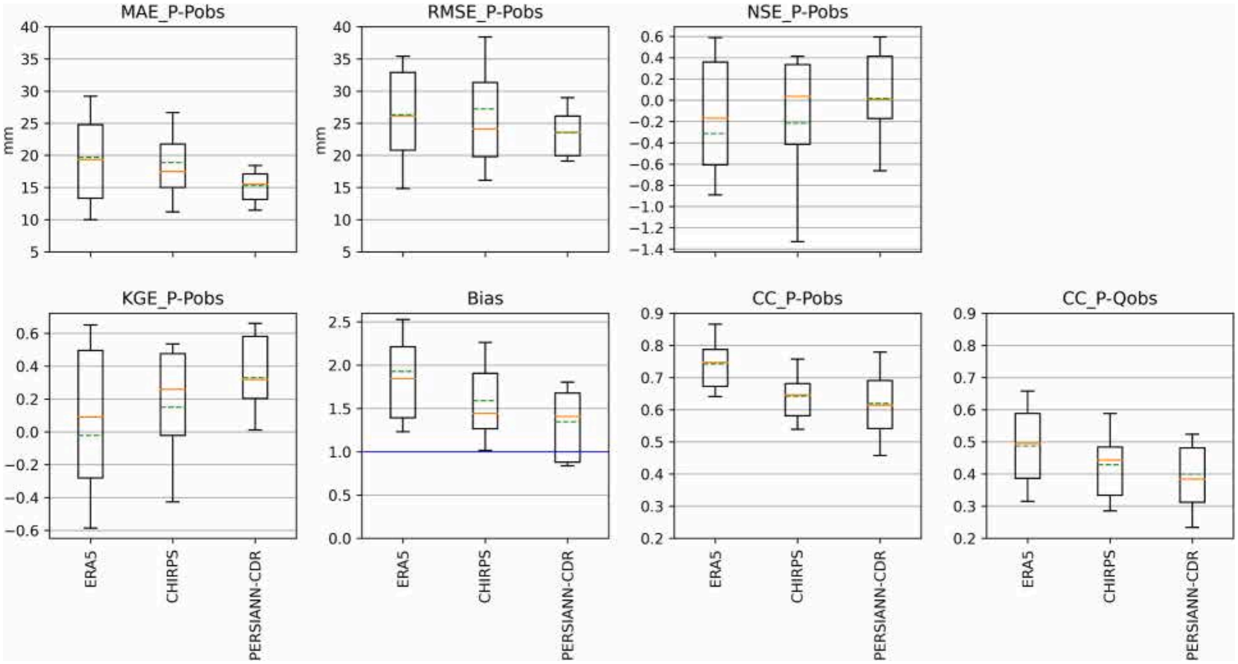

Fig. D.1. Performances of the precipitation products ERA5, CHIRPS, and PERSIANN-CDR with respect to the observed rainfall data, and their correlation with the observed runoff data, over the study catchments during the period of analysis.

Appendix E. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.ejrh.2024.101899.

References

A Lane, R., Coxon, G., E Freer, J., Wagener, T., J Johnes, P., Bloomfield, P., J, et al., 2019. Benchmarking the predictive capability of hydrological models for river flow and flood peak predictions across over 1000 catchments in Great Britain. Hydrol. Earth Syst. Sci. *23* (10), 4011–4032. https://doi.org/10.5194/HESS-23-4011-2019.

Addor, N., Do, H.X., Alvarez-Garreton, C., Coxon, G., Fowler, K., Mendoza, P.A., 2019. Large-sample hydrology: recent progress, guidelines for new datasets and grand challenges. Hydrol. Sci. J. 65 (5), 712–725. https://doi.org/10.1080/02626667.2019.1683182.

Adla, S., Tripathi, S., Disse, M., 2019. Can we calibrate a daily time-step hydrological model using monthly time-step discharge data?, 2019, Vol. 11, 1750 Water *11* (9), 1750. https://doi.org/10.3390/W11091750.

Ahmed, M.I., Stadnyk, T., Pietroniro, A., Awoye, H., Bajracharya, A., Mai, J., et al., 2023. Learning from hydrological models' challenges: a case study from the Nelson basin model intercomparison project. J. Hydrol. *623*, 129820 https://doi.org/10.1016/J.JHYDROL.2023.129820.

Ait Dhmane, L., Moustadraf, J., Rachdane, M., Saidi, M.E., Benjmel, K., Amraoui, F., et al., 2023. Spatiotemporal assessment and correction of gridded precipitation products in North Western Morocco, 2023, Vol. 14, 1239 Atmosphere *14* (8), 1239. https://doi.org/10.3390/ATMOS14081239.

Andréassian, V., Hall, A., Chahinian, N., Schaake, J., 2006. Introduction and synthesis: why should hydrologists work on a large number of basin data sets? Large Sample Basin Exp. Hydrol. Parametr.: Results Models Parameter Exp., IAHS Publ. *307*, 1–5. 〈https://hal.inrae.fr/hal-02588687〉.

Anscombe, F.J., 1973. Graphs in statistical analysis. Am. Stat. 27 (1), 17–21. https://doi.org/10.1080/00031305.1973.10478966.

Arsenault, R., Essou, G.R.C., Brissette, F.P., 2016. Improving hydrological model simulations with combined multi-input and multimodel averaging frameworks. J. Hydrol. Eng. *22* (4), 04016066 https://doi.org/10.1061/(ASCE)HE.1943-5584.0001489.

Ashouri, H., Hsu, K.L., Sorooshian, S., Braithwaite, D.K., Knapp, K.R., Cecil, L.D., et al., 2015. PERSIANN-CDR: daily precipitation climate data record from multisatellite observations for hydrological and climate studies. Bull. Am. Meteorol. Soc. *96* (1), 69–83. https://doi.org/10.1175/BAMS-D-13-00068.1.

Baba, M.W., Boudhar, A., Gascoin, S., Hanich, L., Marchane, A., Chehbouni, A., 2021. Assessment of MERRA-2 and ERA5 to model the snow water equivalent in the high Atlas (1981–2019), 2021, Vol. 13, 890 Water *13* (7), 890. https://doi.org/10.3390/W13070890.

Bai, P., Liu, X., Liang, K., Liu, C., 2015. Comparison of performance of twelve monthly water balance models in different climatic catchments of China. J. Hydrol. *529*, 1030–1040. https://doi.org/10.1016/J.JHYDROL.2015.09.015.

Beck, H.E., Zimmermann, N.E., McVicar, T.R., Vergopolan, N., Berg, A., Wood, E.F., 2018. Present and future Köppen-Geiger climate classification maps at 1-km resolution. *2018 5:1* Sci. Data *5* (1), 1–12. https://doi.org/10.1038/sdata.2018.214.

Benkirane, M., Amazirh, A., Laftouhi, N.E., Khabba, S., Chehbouni, A., 2023. Assessment of GPM satellite precipitation performance after bias correction, for hydrological modeling in a semi-arid watershed (High Atlas Mountain, Morocco), 2023, Vol. 14, 794 Atmosphere *14* (5), 794. https://doi.org/10.3390/ATMOS14050794.

Bisselink, B., Zambrano-Bigiarini, M., Burek, P., de Roo, A., 2016. Assessing the role of uncertain precipitation estimates on the robustness of hydrological model parameters under highly variable climate conditions. J. Hydrol. Reg. Stud. *8*, 112–129. https://doi.org/10.1016/J.EJRH.2016.09.003.

Bouamri, H., Kinnard, C., Boudhar, A., Gascoin, S., Hanich, L., Chehbouni, A., 2021. MODIS does not capture the spatial heterogeneity of snow cover induced by solar radiation. Front. Earth Sci. 9, 640250 https://doi.org/10.3389/feart.2021.640250.

Boudhar, A., Hanich, L., Boulet, G., Duchemin, B., Berjamy, B., Chehbouni, A., 2009. Evaluation of the snowmelt runoff model in the Moroccan high atlas mountains using two snow-cover estimates. Hydrol. Sci. J. *54* (6), 1094–1113. https://doi.org/10.1623/HYSJ.54.6.1094.

Bouizrou, I., Bouadila, A., Aqnouy, M., Gourfi, A., 2023. Assessment of remotely sensed precipitation products for climatic and hydrological studies in arid to semi-arid data-scarce region, central-western Morocco. Remote Sens. Appl. Soc. Environ. *30*, 100976 https://doi.org/10.1016/j.rsase.2023.100976.
Bousbaa, M., Htitiou, A., Boudhar, A., Eljabiri, Y., Elyoussfi, H., Bouamri, H., et al., 2022. High-resolution monitoring of the snow cover on the Moroccan Atlas through the spatio-temporal fusion of landsat and sentinel-2 images, 2022, Vol. 14, 5814 Remote Sens. *14* (22), 5814. https://doi.org/10.3390/RS14225814.
Brocca, L., Melone, F., Moramarco, T., 2011. Distributed rainfall-runoff modelling for flood frequency estimation and flood forecasting. Hydrol. Process. *25* (18), 2801–2813. https://doi.org/10.1002/HYP.8042.
Budyko, M.I., 1958. The heat balance of the earth's surface. Sov. Geogr. *2* (4), 3–13. https://doi.org/10.1080/00385417.1961.10770761.
Camici, S., Ciabatta, L., Massari, C., Brocca, L., 2018. How reliable are satellite precipitation estimates for driving hydrological models: a verification study over the Mediterranean area. J. Hydrol. *563*, 950–961. https://doi.org/10.1016/J.JHYDROL.2018.06.067.
Camici, Stefania, Massari, C., Ciabatta, L., Marchesini, I., Brocca, L., 2020. Which rainfall score is more informative about the performance in river discharge simulation? A comprehensive assessment on 1318 basins over Europe. Hydrol. Earth Syst. Sci. *24* (10), 4869–4885. https://doi.org/10.5194/HESS-24-4869-2020.
Chlumsky, R., Mai, J., Craig, J.R., Tolson, B.A., 2021. Simultaneous calibration of hydrologic model structure and parameters using a blended model. Water Resour. Res. *57* (5), e2020WR029229 https://doi.org/10.1029/2020WR029229.
Coron, L., Andréassian, V., Perrin, C., Lerat, J., Vaze, J., Bourqui, M., Hendrickx, F., 2012. Crash testing hydrological models in contrasted climate conditions: an experiment on 216 Australian catchments. Water Resour. Res. *48* (5), 5552. https://doi.org/10.1029/2011WR011721.
Dakhlaoui, H., Ruelland, D., Tramblay, Y., Bargaoui, Z., 2017. Evaluating the robustness of conceptual rainfall-runoff models under climate variability in northern Tunisia. J. Hydrol. *550*, 201–217. https://doi.org/10.1016/J.JHYDROL.2017.04.032.
Darbandsari, P., Coulibaly, P., 2020. Inter-comparison of lumped hydrological models in data-scarce watersheds using different precipitation forcing data sets: case study of Northern Ontario, Canada. J. Hydrol.: Reg. Stud. *31*, 100730 https://doi.org/10.1016/J.EJRH.2020.100730.
De la Fuente, L.A., Gupta, H.V., Condon, L.E., 2023. Toward a multi-representational approach to prediction and understanding, in support of discovery in hydrology. Water Resour. Res. *59* (1), e2021WR031548 https://doi.org/10.1029/2021WR031548.
Driouech, F., Mahé, G., Déqué, M., Dieulin, C., El Heirech, T., Milano, M., et al., 2010. Evaluation d'impacts potentiels de changements climatiques sur l'hydrologie du bassin versant de la Moulouya au Maroc. IAHS-AISH Publ. *340*, 561–567.
Dutta, P., Sarma, A.K., 2021. Hydrological modeling as a tool for water resources management of the data-scarce Brahmaputra basin. J. Water Clim. Change *12* (1), 152–165. https://doi.org/10.2166/WCC.2020.186.
El Khalki, E.M., Tramblay, Y., Hanich, L., Marchane, A., Boudhar, A., Hakkani, B., 2021. Climate change impacts on surface water resources in the Oued El Abid basin, Morocco. Hydrol. Sci. J. *66* (15), 2132–2145. https://doi.org/10.1080/02626667.2021.1982137.
El Khalki, E.M., Tramblay, Y., Saidi, M.E., Marchane, A., Chehbouni, A., 2023. Hydrological assessment of different satellite precipitation products in semi-arid basins in Morocco. Front. Water *5*, 1243251. https://doi.org/10.3389/FRWA.2023.1243251.
Elair, C., Rkha Chaham, K., Hadri, A., 2023. Assessment of drought variability in the Marrakech-Safi region (Morocco) at different time scales using GIS and remote sensing. Water Supply. https://doi.org/10.2166/WS.2023.283.
Er-Raki, S., Chehbouni, A., Khabba, S., Simonneaux, V., Jarlan, L., Ouldbba, A., et al., 2010. Assessment of reference evapotranspiration methods in semi-arid regions: Can weather forecast data be used as alternate of ground meteorological parameters? J. Arid Environ. *74* (12), 1587–1596. https://doi.org/10.1016/J.JARIDENV.2010.07.002.
Essou, G.R.C., Arsenault, R., Brissette, F.P., 2016. Comparison of climate datasets for lumped hydrological modeling over the continental United States. J. Hydrol. *537*, 334–345. https://doi.org/10.1016/J.JHYDROL.2016.03.063.
Farr, T.G., Rosen, P.A., Caro, E., Crippen, R., Duren, R., Hensley, S., et al., 2007. The shuttle radar topography mission. Rev. Geophys. *45* (2), 2004. https://doi.org/10.1029/2005RG000183.
Fathi, M.M., Awadallah, A.G., Aldahshoory, W., 2023. An improved monthly water balance GR2M model with a seasonally variable parameter. J. Hydrol. *617*, 129127 https://doi.org/10.1016/J.JHYDROL.2023.129127.
Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., et al., 2015. The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes. *2015 2:1* Sci. Data *2* (1), 1–21. https://doi.org/10.1038/sdata.2015.66.
Gan, T.Y., Dlamini, E.M., Fana Biftu, G., 1997. Effects of model complexity and structure, data quality, and objective functions on hydrologic modeling. J. Hydrol. *192*, 81–103 https://doi.org/10.1016/S0022-1694(96)03114-9.
Gong, H., Pan, Y., Xu, Y., 2012. Spatio-temporal variation of groundwater recharge in response to variability in precipitation, land use and soil in Yanqing Basin, Beijing, China. Hydrogeol. J. *20* (7), 1331–1340. https://doi.org/10.1007/s10040-012-0883-x.
Gowda, P.H., Mulla, D.J., Desmond, E.D., Ward, A.D., Moriasi, D.N., 2012. ADAPT: model use, calibration, and validation. Trans. ASABE *55* (4), 1345–1352 https://doi.org/10.13031/2013.42246.
Gupta, H.V., Kling, H., Yilmaz, K.K., Martinez, G.F., 2009. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. J. Hydrol. *377* (1–2), 80–91. https://doi.org/10.1016/J.JHYDROL.2009.08.003.
Gupta, H.V., Perrin, C., Blöschl, G., Montanari, A., Kumar, R., Clark, M., Andréassian, V., 2014. Large-sample hydrology: a need to balance depth with breadth. Hydrol. Earth Syst. Sci. *18*, 463–477. https://doi.org/10.5194/hess-18-463-2014.
Hadri, A., Saidi, M.E.M., Saouabe, T., El Fels, A.E.A., 2021. Temporal trends in extreme temperature and precipitation events in an arid area: case of Chichaoua Mejjate region (Morocco). J. Water Clim. Change *12* (3), 895–915. https://doi.org/10.2166/WCC.2020.234.
Hanich, L., Chehbouni, A., Gascoin, S., Boudhar, A., Jarlan, L., Tramblay, Y., et al., 2022. Snow hydrology in the Moroccan Atlas Mountains. J. Hydrol.: Reg. Stud. *42*, 101101 https://doi.org/10.1016/J.EJRH.2022.101101.
Hapuarachchi, H.A.P., Bari, M.A., Kabir, A., Hasan, M.M., Woldemeskel, F.M., Gamage, N., et al., 2022. Development of a national 7-day ensemble streamflow forecasting service for Australia. Hydrol. Earth Syst. Sci. *26* (18), 4801–4821. https://doi.org/10.5194/HESS-26-4801-2022.
Hargreaves, G.H., Samani, Z.A., 1985. Reference Crop Evapotranspiration from Temperature. Appl. Eng. Agric. 1 (2), 96–99. https://doi.org/10.13031/2013.26773.
Hengl, T., Miller, M.A.E., Križan, J., Shepherd, K.D., Sila, A., Kilibarda, M., et al., 2021. African soil properties and nutrients mapped at 30 m spatial resolution using two-scale ensemble machine learning. *2021 11:1 Sci. Rep. 11* (1), 1–18. https://doi.org/10.1038/s41598-021-85639-y.
Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., et al., 2020. The ERA5 global reanalysis. Q. J. R. Meteorol. Soc. *146* (730), 1999–2049. https://doi.org/10.1002/QJ.3803.
Hrour, Y., Fovet, O., Lacombe, G., Rousseau-Gueutin, P., Sebari, K., Pichelin, P., Thomas, Z., 2023. A framework to assess future water-resource under climate change in northern Morocco using hydro-climatic modelling and water-withdrawal scenarios. J. Hydrol.: Reg. Stud. *48*, 101465 https://doi.org/10.1016/J.EJRH.2023.101465.
Jain, V.K., Pandey, R.P., Jain, M.K., Byun, H.R., 2015. Comparison of drought indices for appraisal of drought characteristics in the Ken River Basin. Weather Clim. Extrem. *8*, 1–11. https://doi.org/10.1016/J.WACE.2015.05.002.
Jaiswal, R.K., Ali, S., Bharti, B., 2020. Comparative evaluation of conceptual and physical rainfall–runoff models. Appl. Water Sci. 10 (1), 1–14. https://doi.org/10.1007/S13201-019-1122-6.
Jakeman, A.J., Hornberger, G.M., 1993. How much complexity is warranted in a rainfall-runoff model? Water Resour. Res. *29* (8), 2637–2649. https://doi.org/10.1029/93WR00877.
Jayathilake, D.I., Smith, T., 2021. Assessing the impact of PET estimation methods on hydrologic model performance. Hydrol. Res. *52* (2), 373–388. https://doi.org/10.2166/NH.2020.066.
Kim, H.S., Lee, S., 2014. Assessment of the adequacy of the regional relationships between catchment attributes and catchment response dynamics, calibrated by a multi-objective approach. Hydrol. Process. *28* (13), 4023–4041. https://doi.org/10.1002/HYP.9942.
Knoben, W.J.M., Freer, J.E., Peel, M.C., Fowler, K.J.A., Woods, R.A., 2020. A Brief Analysis of Conceptual Model Structure Uncertainty Using 36 Models and 559 Catchments. Water Resour. Res. *56* (9), e2019WR025975 https://doi.org/10.1029/2019WR025975.

Kouakou, C., Paturel, J.-E., Satgé, F., Tramblay, Y., Defrance, D., Rouché, N., 2023. Comparison of gridded precipitation estimates for regional hydrological modeling in West and Central Africa. J. Hydrol.: Reg. Stud. *47*, 2214–5818. https://doi.org/10.1016/j.ejrh.2023.101409.

Li, C.Z., Zhang, L., Wang, H., Zhang, Y.Q., Yu, F.L., Yan, D.H., 2012. The transferability of hydrological models under nonstationary climatic conditions. Hydrol. Earth Syst. Sci. *16* (4), 1239–1254. https://doi.org/10.5194/HESS-16-1239-2012.

López-Moreno, J.I., Granados, I., Ceballos-Barbancho, A., Morán-Tejeda, E., Revuelto, J., Alonso-González, E., et al., 2023. The signal of snowmelt in streamflow and stable water isotopes in a high mountain catchment in Central Spain. J. Hydrol.: Reg. Stud. *46*, 101356 https://doi.org/10.1016/J.EJRH.2023.101356.

Mahmoudi, P., Rigi, A., Miri Kamak, M., 2019. A comparative study of precipitation-based drought indices with the aim of selecting the best index for drought monitoring in Iran. Theor. Appl. Climatol. 137 (3–4), 3123–3138. https://doi.org/10.1007/S00704-019-02778-Z.

Marchane, A., Jarlan, L., Hanich, L., Boudhar, A., Gascoin, S., Tavernier, A., et al., 2015. Assessment of daily MODIS snow cover products to monitor snow cover dynamics over the Moroccan Atlas mountain range. Remote Sens. Environ. *160*, 72–86. https://doi.org/10.1016/J.RSE.2015.01.002.

Massmann, C., 2020. Identification of factors influencing hydrologic model performance using a top-down approach in a large number of U.S. catchments. Hydrol. Process. *34* (1), 4–20. https://doi.org/10.1002/HYP.13566.

Mathevet, T., Gupta, H., Perrin, C., Andréassian, V., Le Moine, N., 2020. Assessing the performance and robustness of two conceptual rainfall-runoff models on a worldwide sample of watersheds. J. Hydrol. *585*, 124698 https://doi.org/10.1016/J.JHYDROL.2020.124698.

McMillan, H.K., Booker, D.J., Cattoën, C., 2016. Validation of a national hydrological model. J. Hydrol. *541*, 800–815. https://doi.org/10.1016/J.JHYDROL.2016.07.043.

Merz, R., Parajka, J., Blöschl, G., 2011. Time stability of catchment model parameters: Implications for climate impact analyses. Water Resour. Res. *47* (2), 2531. https://doi.org/10.1029/2010WR009505.

Motavita, D.F., Chow, R., Guthke, A., Nowak, W., 2019. The comprehensive differential split-sample test: a stress-test for hydrological model robustness under climate variability. J. Hydrol. *573*, 501–515. https://doi.org/10.1016/J.JHYDROL.2019.03.054.

Mouelhi, S., Michel, C., Perrin, C., Andréassian, V., 2006. Stepwise development of a two-parameter monthly water balance model. J. Hydrol. *318* (1–4), 200–214. https://doi.org/10.1016/j.jhydrol.2005.06.014.

Muñoz-Castro, E., Mendoza, P.A., Vásquez, N., Vargas, X., 2023. Exploring parameter (dis)agreement due to calibration metric selection in conceptual rainfall–runoff models. Hydrol. Sci. J. *68* (12), 1754–1768. https://doi.org/10.1080/02626667.2023.2231434.

Najmi, A., Igmoullan, B., Namous, M., El Bouazzaoui, I., Brahim, Y.A., El Khalki, E.M., Saidi, M.E.M., 2023. Evaluation of PERSIANN-CCS-CDR, ERA5, and SM2RAIN-ASCAT rainfall products for rainfall and drought assessment in a semi-arid watershed, Morocco. J. Water Clim. Change *14* (5), 1569–1584. https://doi.org/10.2166/WCC.2023.461.

Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models part I — A discussion of principles. J. Hydrol. *10* (3), 282–290. https://doi.org/10.1016/0022-1694(70)90255-6.

Orth, R., Staudinger, M., Seneviratne, S.I., Seibert, J., Zappa, M., 2015. Does model performance improve with complexity? A case studywith three hydrological models. J. Hydrol. *523*, 147–159. https://doi.org/10.1016/J.JHYDROL.2015.01.044.

Ouatiki, H., Boudhar, A., Chehbouni, A., 2023. Accuracy assessment and bias correction of remote sensing–based rainfall products over semiarid watersheds. Theor. Appl. Climatol. 154 (3–4), 763–780. https://doi.org/10.1007/S00704-023-04586-Y.

Ouatiki, H., Boudhar, A., Tramblay, Y., Jarlan, L., Benabdelouhab, T., Hanich, L., et al., 2017. Evaluation of TRMM 3B42 V7 Rainfall Product over the Oum Er Rbia Watershed in Morocco, 2017, Vol. 5, 1 Climate *5* (1), 1. https://doi.org/10.3390/CLI5010001.

Oudin, L., Michel, C., Anctil, F., 2005. Which potential evapotranspiration input for a lumped rainfall-runoff model? Part 1 - Can rainfall-runoff models effectively handle detailed potential evapotranspiration inputs? J. Hydrol. *303* (1–4), 275–289. https://doi.org/10.1016/j.jhydrol.2004.08.025.

Pelletier, A., Andréassian, V., 2022. On constraining a lumped hydrological model with both piezometry and streamflow: results of a large sample evaluation. Hydrol. Earth Syst. Sci. *26*, 2733–2758. https://doi.org/10.5194/hess-26-2733-2022.

Perrin, C., Michel, C., Andréassian, V., 2001. Does a large number of parameters enhance model performance? Comparative assessment of common catchment model structures on 429 catchments. J. Hydrol. *242* (3–4), 275–301. https://doi.org/10.1016/S0022-1694(00)00393-0.

Pham, L.T., Luo, L., Finley, A., 2021. Evaluation of random forests for short-term daily streamflow forecasting in rainfall-and snowmelt-driven watersheds. Hydrol. Earth Syst. Sci. *25*, 2997–3015. https://doi.org/10.5194/hess-25-2997-2021.

Pimentel, R., Arheimer, B., Crochemore, L., Andersson, J.C.M., Pechlivanidis, I.G., Gustafsson, D., 2023. Which Potential Evapotranspiration Formula to Use in Hydrological Modeling World-Wide? Water Resour. Res. *59* (5), e2022WR033447 https://doi.org/10.1029/2022WR033447.

Poncelet, C., Merz, R., Merz, B., Parajka, J., Oudin, L., Andréassian, V., Perrin, C., 2017. Process-based interpretation of conceptual hydrological model performance using a multinational catchment set. Water Resour. Res. *53* (8), 7247–7268. https://doi.org/10.1002/2016WR019991.

Rachdane, M., Khalki, E.M., El, Saidi, M.E., Nehmadou, M., Ahbari, A., Tramblay, Y., 2022. Comparison of High-Resolution Satellite Precipitation Products in Sub-Saharan Morocco. Water (Switz.) 14 (20), 3336. https://doi.org/10.3390/W14203336.

Rachidi, S., El Mazoudi, E.H., El Alami, J., Jadoud, M., Er-Raki, S., 2023. Assessment and Comparison of Satellite-Based Rainfall Products: Validation by Hydrological Modeling Using ANN in a Semi-Arid Zone, 2023, Vol. 15, 1997 Water *15* (11), 1997. https://doi.org/10.3390/W15111997.

Rajurkar, M.P., Kothyari, U.C., Chaube, U.C., 2004. Modeling of the daily rainfall-runoff relationship with artificial neural network. J. Hydrol. *285* (1–4), 96–113. https://doi.org/10.1016/J.JHYDROL.2003.08.011.

Rhoujjati, N., Ait Brahim, Y., Hanich, L., Rhoujjati, A., Rafik, A., Ouatiki, H., et al., 2023. Snowpack and groundwater recharge in the Atlas mountains: New evidence and key drivers. J. Hydrol.: Reg. Stud. *49*, 101520 https://doi.org/10.1016/J.EJRH.2023.101520.

Salih, W., Chehbouni, A., Epule, T.E., 2022. Evaluation of the Performance of Multi-Source Satellite Products in Simulating Observed Precipitation over the Tensift Basin in Morocco. Remote Sens. 14 (5), 1171. https://doi.org/10.3390/RS14051171.

Salih, W., Epule, T.E., EL Khalki, E.M., Ouatiki, H., Erraki, S., Achli, S., Chehbouni, A., 2023. A comprehensive assessment of satellite precipitation products over a semi-arid region: focus on extreme events. *Nat. Hazards 2023*, 1–29. https://doi.org/10.1007/S11069-023-06317-Y.

Saouabe, T., El Khalki, E.M., Saidi, M.E.M., Najmi, A., Hadri, A., Rachidi, S., et al., 2020. Evaluation of the GPM-IMERG Precipitation Product for Flood Modeling in a Semi-Arid Mountainous Basin in Morocco, 2020, Vol. 12, 2516 Water *12* (9), 2516. https://doi.org/10.3390/W12092516.

Saouabe, T., Naceur, K.A., Khalki, E.M., El, Hadri, A., Saidi, M.E., 2022. GPM-IMERG product: a new way to assess the climate change impact on water resources in a Moroccan semi-arid basin. J. Water Clim. Change 13 (7), 2559–2576. https://doi.org/10.2166/wcc.2022.403.

Schoups, G., Van De Giesen, N.C., Savenije, H.H.G., 2008. Model complexity control for hydrologic prediction. Water Resour. Res. *44* (12) https://doi.org/10.1029/2008WR006836.

Seibert, J., Staudinger, M., van Meerveld, H.J. (Ilja), 2019. Validation and Over-Parameterization—Experiences from Hydrological Modeling. Computer Simulation Validation. Springer, Cham, pp. 811–834. https://doi.org/10.1007/978-3-319-70766-2_33.

Singh, V., Muñoz-Arriola, F., 2021. Improvements in sub-catchment fractional snowpack and snowmelt parameterizations and hydrologic modeling for climate change assessments in the western Himalayas. Hydrology 8 (4), 179. https://doi.org/10.3390/HYDROLOGY8040179.

Stephens, C.M., Pham, H.T., Marshall, L.A., Johnson, F.M., 2022. Which Rainfall Errors Can Hydrologic Models Handle? Implications for Using Satellite-Derived Products in Sparsely Gauged Catchments. Water Resour. Res. *58* (8), e2020WR029331 https://doi.org/10.1029/2020WR029331.

Storn, R., Price, K., 1997. Differential Evolution - A Simple and Efficient Heuristic for Global Optimization over Continuous Spaces. J. Glob. Optim. 11 (4), 341–359. https://doi.org/10.1023/A:1008202821328.

Sulla-Menashe, D., & Friedl, M.A. (2022). [dataset] MODIS/Terra+Aqua Land Cover Type Yearly L3 Global 500m SIN Grid V061. https://doi.org/10.5067/MODIS/MCD12Q1.

Tarek, M., Arsenault, R., Brissette, F., Martel, J.L., 2021. Daily streamflow prediction in ungauged basins: an analysis of common regionalization methods over the African continent. Hydrol. Sci. J. *66* (11), 1695–1711. https://doi.org/10.1080/02626667.2021.1948046.

Tarek, M., Brissette, F.P., Arsenault, R., 2020. Comparison of gridded datasets for the simulation of streamflow in Africa. IOP Conf. Ser.: Mater. Sci. Eng. *974* (1), 012001 https://doi.org/10.1088/1757-899X/974/1/012001.

Teegavarapu, R.S.V., Aly, A., Pathak, C.S., Ahlquist, J., Fuelberg, H., Hood, J., 2018. Infilling missing precipitation records using variants of spatial interpolation and data-driven methods: use of optimal weighting parameters and nearest neighbour-based corrections. Int. J. Climatol. *38* (2), 776–793. https://doi.org/10.1002/JOC.5209.

Teegavarapu, R.S.V., Chandramouli, V., 2005. Improved weighting methods, deterministic and stochastic data-driven models for estimation of missing precipitation records. J. Hydrol. *312* (1–4), 191–206. https://doi.org/10.1016/J.JHYDROL.2005.02.015.

Th van Genuchten, by M., Leij, F.J., Yates, S.R., Williams, J.R., & Kerr, R.S. (1991). The RETC Code for Quantifying the Hydraulic Functions of Unsaturated Soils. *U.S. Environmental Protection Agency*.

Thomas Jr. Harold A. (1981). Improved methods for national water assessment, water resources contract: WR15249270. *USGS Publications Warehouse*. doi.org/10.3133/70046351.

Tramblay, Y., El Khalki, E.M., Ciabatta, L., Camici, S., Hanich, L., Saidi, M.E.M., et al., 2023. River runoff estimation with satellite rainfall in Morocco. Hydrol. Sci. J. https://doi.org/10.1080/02626667.2023.2171295.

Tramblay, Y., Thiemig, V., Dezetter, A., Hanich, L., 2016. Evaluation of satellite-based rainfall products for hydrological modelling in Morocco. Hydrol. Sci. J. *61* (14), 2509–2519. https://doi.org/10.1080/02626667.2016.1154149.

Van Esse, W.R., Perrin, C., Booij, M.J., Augustijn, D.C.M., Fenicia, F., Kavetski, D., Lobligeois, F., 2013. The influence of conceptual model structure on model performance: A comparative study for 237 French catchments. Hydrol. Earth Syst. Sci. *17* (10), 4227–4239. https://doi.org/10.5194/hess-17-4227-2013.

Vandewiele, G.L., Ni-Lar-Win, 2009. Monthly water balance models for 55 basins in 10 countries. Hydrol. Sci. J. *43* (5), 687–699. https://doi.org/10.1080/02626669809492166.

Vandewiele, G., Xu, C.-Y., 1992. Methodology and comparative study of monthly water balance models in Belgium, China and Burma. J. Hydrol. *134*, 315–347. https://doi.org/10.1016/0022-1694(92)90041-S.

Vaze, J., Post, D.A., Chiew, F.H.S., Perraud, J.M., Viney, N.R., Teng, J., 2010. Climate non-stationarity – Validity of calibrated rainfall–runoff models for use in climate change studies. J. Hydrol. *394* (3–4), 447–457. https://doi.org/10.1016/J.JHYDROL.2010.09.018.

Wan, Y., Chen, J., Xu, C.Y., Xie, P., Qi, W., Li, D., Zhang, S., 2021. Performance dependence of multi-model combination methods on hydrological model calibration strategy and ensemble size. J. Hydrol. *603*, 127065 https://doi.org/10.1016/J.JHYDROL.2021.127065.

Wang, Q.J., Pagano, T.C., Zhou, S.L., Hapuarachchi, H.A.P., Zhang, L., Robertson, D.E., 2011. Monthly versus daily water balance models in simulating monthly runoff. J. Hydrol. *404* (3–4), 166–175. https://doi.org/10.1016/J.JHYDROL.2011.04.027.

Wang, G.Q., Zhang, J.Y., Jin, J.L., Liu, Y.L., He, R.M., Bao, Z.X., et al., 2014. Regional calibration of a water balance model for estimating stream flow in ungauged areas of the Yellow River Basin. Quat. Int. *336*, 65–72. https://doi.org/10.1016/J.QUAINT.2013.08.051.

Wang, J., Zhuo, L., Han, D., Liu, Y., Rico-Ramirez, M.A., 2023. Hydrological Model Adaptability to Rainfall Inputs of Varied Quality. Water Resour. Res. *59* (2), e2022WR032484 https://doi.org/10.1029/2022WR032484.

Xiong, L., Guo, S., 1999. A two-parameter monthly water balance model and its application. J. Hydrol. *216* (1–2), 111–123. https://doi.org/10.1016/S0022-1694(98)00297-2.

Xu, C.Y., Singh, V.P., 1998. A Review on Monthly Water Balance Models for Water Resources Investigations. Water Resour. Manag. 12 (1), 20–50. https://doi.org/10.1023/A:1007916816469.

Xu, Z., Wu, Z., Shao, Q., He, H., Guo, X., 2022. A two-step calibration framework for hydrological parameter regionalization based on streamflow and remote sensing evapotranspiration. J. Hydrol. *613*, 128320 https://doi.org/10.1016/J.JHYDROL.2022.128320.

Ye, W., Bates, B.C., Viney, N.R., Sivapalan, M., Jakeman, A.J., 1997. Performance of conceptual rainfall-runoff models in low-yielding ephemeral catchments. Water Resour. Res. *33* (1), 153–166. https://doi.org/10.1029/96WR02840.

Zamoum, S., Souag-Gamane, D., 2019. Monthly streamflow estimation in ungauged catchments of northern Algeria using regionalization of conceptual model parameters. Arab. J. Geosci. 12 (11), 1–14. https://doi.org/10.1007/S12517-019-4487-9.

Zhang, K., Luhar, M., Brunner, M.I., Parolari, A.J., 2023. Streamflow prediction in poorly gauged watersheds in the United States through data-driven sparse sensing. Water Resour. Res. *59* (4), e2022WR034092 https://doi.org/10.1029/2022WR034092.

Zhang, L., Potter, N., Hickel, K., Zhang, Y., Shao, Q., 2008. Water balance modeling over variable time scales based on the Budyko framework – model development and testing. J. Hydrol. *360* (1–4), 117–131. https://doi.org/10.1016/J.JHYDROL.2008.07.021.