Improving Streamflow Predictions in the Arid Southwestern United States Through Understanding of Baseflow Generation Mechanisms

Mohammad A. Farmani[1], Ahmad Tavakoly[2,3], Ali Behrangi[1,4], Yuan Qiu[1,5], Aniket Gupta[1], Muhammad Jawad[1], Hossein Yousefi Sohi[1], Xueyan Zhang[1], Matthew Geheran[1], Guo-Yue Niu[1]

[1]Department of Hydrology and Atmospheric Sciences, University of Arizona, Tucson, AZ, USA,
[2]US Army Engineer Research and Development Center, Coastal and Hydraulics Laboratory, Vicksburg, MS, USA,
[3]Earth System Science Interdisciplinary Center, University of Maryland, College Park, MD, USA
[4]Department of Geosciences, University of Arizona, Tucson, AZ, USA,
[5]Center for Hydrologic Innovations, School of Sustainability, Arizona State University, Tempe, AZ, USA.

Corresponding author: Mohammad Farmani, email: farmani@arizona.edu

Abstract

Understanding factors controlling baseflow (or groundwater discharge) is critical for improving streamflow prediction skills in the arid southwest US. We used a version of Noah-MP with newly-advanced hydrology features and the Routing Application for Parallel computation of Discharge (RAPID) to investigate the impacts of uncertainties in representations of hydrological processes, soil hydraulic parameters, and precipitation data on baseflow production and streamflow prediction skill. We conducted model experiments by combining different options of hydrological processes, hydraulic parameters, and precipitation datasets in the southwest US. These experiments were driven by three gridded precipitation products: the North American Land Data Assimilation System (NLDAS-2), the Integrated Multi-satellite Retrievals for GPM (IMERG) Final, and the NOAA Analysis of Record for Calibration (AORC). RAPID was then used to route Noah-MP modeled surface and subsurface runoff to predict daily streamflow at 390 USGS gauges. We evaluated the modeled ratio of baseflow to total streamflow (or baseflow index, BFI) against those derived from the USGS streamflow. Our results suggest that 1) soil water retention curve model plays a dominant role, with the Van-Genuchten hydraulic scheme reducing the overestimated BFI produced by the Brooks-Corey (also used by the National Water Model, NWM), 2) hydraulic parameters strongly affect streamflow prediction, a machine learning-based dataset captures the USGS BFI, showing a better performance than the optimized NWM by a median KGE of 21%, and 3) the ponding depth threshold that increases infiltration is preferred. Overall, most of our models with the advanced hydrology show a better performance in modeling BFI and thus a better skill in streamflow predictions than the optimized NWM in the dry southwestern river basins. These findings can guide future studies in selecting reliable schemes and datasets (before calibration) to achieve better streamflow predictions as well as water resource projections.

1 Introduction

In arid regions, accurate streamflow prediction presents a significant challenge due to complexities in baseflow generation, which are influenced by highly variable precipitation in time and space, as well as heterogeneous properties of soil, snow, and vegetation across complex terrain. The complexity of the hydrological processes in these areas complicates efforts to estimate water availability, which is essential for effective water resource management, agricultural planning, and disaster preparedness. (Thomas 1994, Poff, Allan et al. 1997, Su, Lettenmaier et al. 2024). Large-scale hydrological models such as the National Water Model (NWM) are employed to predict streamflow across the nation. The NWM system integrates various data sources, including meteorological inputs and land surface characteristics, to provide comprehensive hydrological insights. Similarly, the performance of other Land surface models such as VIC (Lohmann, Nolte-Holube et al. 1996), ParFlow-CONUS (Tijerina-Kreuzer, Condon et al. 2021), and SAC-SMA (Burnash 1995) in arid regions is often limited, as they struggle to accurately simulate the hydrological dynamics of the dry environments (Wheater and Evans 2009, Sivapalan, Savenije et al. 2012, Blöschl, Sivapalan et al. 2013, Ghimire, Hansen et al. 2023, Johnson, Fang et al. 2023, Towler, Foks et al. 2023).

For instance, Salas et al. (2017) evaluated an uncalibrated version of Weather Research and Forecasting Model Hydrological (WRF-Hydro) and noted relatively weaker performance in the arid regions of the Texas Gulf Coast basin (Salas, Somos-Valenzuela et al. 2017). Similarly, Lin et al. (2018) utilized a WRF-Hydro-RAPID framework to simulate streamflow in Texas and identified a significant positive bias over dry regions, attributing this to overprediction of baseflow and surface runoff (Lin, Rajib et al. 2018).

Hansen et al. (2019) identified two significant issues with the NWM streamflow hindcasts over the Colorado River basin. First, they observed that the NWM tends to underestimate the frequency of low flows. Second, the model inaccurately identifies locations as experiencing low flow where it does not actually occur, while failing to detect low flow in locations where it is present (Hansen, Shiva et al. 2019). Towler et al. (2023) reported similar issues in both NWM and the National Hydrologic Model (NHM) at gauges in the central and southwestern United States. This tendency of underperformance in the arid southwestern U.S. was also observed in models such as VIC, ParFlow-CONUS, and SAC-SMA (Newman, Clark et al. 2015, Tijerina-Kreuzer, Condon et al. 2021, Ghimire, Hansen et al. 2023, Towler, Foks et al. 2023). Despite efforts to improve the Noah-MP land surface exchange scheme of NWM, including various calibration attempts, predictions in arid regions remain inadequate (Bass, Rahimi et al. 2023, Su, Lettenmaier et al. 2024). Hence, there is a need for further research to understand the uncertainties in runoff generation processes to improve their performance in streamflow prediction in the southwestern U.S.

Large-scale hydrological models often exhibit uncertainties in representing surface runoff and streamflow generation due to varying approaches for estimating soil hydraulic properties, which are crucial for understanding these processes (Vereecken, Weihermüller et al. 2019). Baseflow plays a significant role in streamflow generation, particularly in arid regions where it sustains water flow during dry periods (Sophocleous 2002). Factors affecting baseflow—and consequently runoff and streamflow—in these regions include soil moisture content, groundwater recharge rates, soil hydraulic properties, and soil structure (Scanlon, Keese et al. 2006). A key concern is the lack of a consistent framework for predicting effective fluxes and infiltration parameters in LSMs. Additionally, the influence of soil structure on hydraulic properties, which affects infiltration rates and baseflow, is often overlooked (Vereecken, Weihermüller et al. 2019). Variations in soil hydraulic schemes significantly impact soil moisture and drainage(Farmani, Behrangi et al. 2024), both essential for sustaining baseflow (van Dijk 2010) — a key parameter that may contribute to the poor performance of large-scale hydrological models like NWM in arid regions. Moreover, while large-scale models often overlook the specific impacts of ponding on soil moisture (Niu, Fang et al. 2024), some studies have shown the influence of ponding depth on soil moisture and water release (Farmani, Behrangi et al. 2024). However, research integrating ponding effects into large-scale models, particularly concerning baseflow in arid regions, remains scant.

Uncertainties in precipitation data—particularly in amount, intensity, and spatial-temporal resolution—significantly affect runoff, groundwater recharge and streamflow generation. Recharge is governed by precipitation characteristics such as duration, magnitude, and intensity (Crosbie, McCallum et al. 2012, Dourte, Shukla et al. 2013, Huang, Wu et al. 2013, Moghisi, Yazdi et al. 2024). While classical theory asserts that low-intensity rainfall over long periods generates the highest fractional recharge (Dourte, Shukla et al. 2013), recent studies indicate that in certain regions, such as East Africa and Australia, extreme rainfall events drive the majority of recharge (Kendy, Gérard-Marchant et al. 2003, Kendy, Zhang et al. 2004, Crosbie, McCallum et al. 2012). Coarse temporal averaging (e.g., monthly or annual) smooths these events, underestimating recharge rates. For example, daily versus yearly data can produce recharge estimates up to nine times higher (Tashie, Mirus et al. 2016, Batalha, Barbosa et al. 2018). This is especially important in areas with distinct wet and dry seasons, where episodic recharge events are essential for maintaining baseflow during dry periods. Furthermore, spatial resolution plays a crucial role, as coarse grid models may fail to capture localized rainfall events, exacerbating uncertainties in recharge and baseflow predictions (Mileham, Taylor et al. 2009).

As baseflow is recognized as the primary source of streamflow in dry regions (van Dijk 2010), we hypothesize that the baseflow generation processes in the hydrological models contribute to the inaccuracies in streamflow predictions (Figure 1). We used a version of Noah-MP, which is used in NWM as the surface exchange and runoff generation scheme in NWM, with a newly-advanced hydrology by implementing the mixed-form Richards equation down to the bedrock alongside the Routing Application for Parallel computation of Discharge (RAPID, David et al., 2011) to understand the major factors affecting the baseflow generation for improved streamflow predictions. We conducted various model experiments to study the sensitivity of baseflow generation to soil hydraulic schemes, soil water retention curve parameter datasets, model physics (single vs. dual permeability schemes), surface ponding thresholds, and precipitation datasets across the southwestern US. RAPID was then used to route Noah-MP modeled surface and subsurface runoff (or groundwater discharge) to predict daily streamflow in 390 USGS gauges from 1980 to 2019. It should be noted that groundwater discharge and subsurface runoff are considered the same in Noah-MP. We evaluated the Noah-MP modeled Baseflow Index (BFI) and NWM BFI against those derived from the USGS streamflow. We aim to provide guidance for future research in selecting the most reliable schemes and datasets to enhance streamflow predictions, particularly for dry regions.

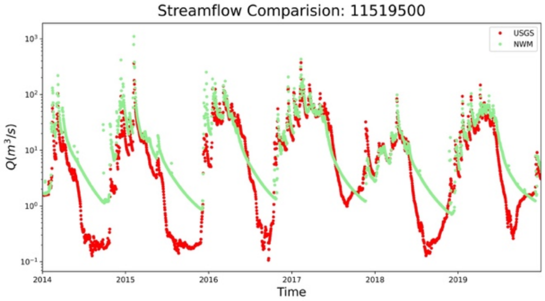

Figure 1 Observed USGS (red) streamflow for the Scott River near Fort Jones, CA – 11519500 station, and NWM v2.1 (green) prediction. NWM shows an overestimation of low flows and a gradual slope on the falling limbs.

BFI measures the percentage of streamflow that comes from baseflow over a long time (van Dijk 2010, Huang, Dong et al. 2021). BFI is commonly used in hydrology to characterize low flow conditions within a catchment, providing insights into the sustained contribution of groundwater and other delayed sources to total streamflow (Seo, Mahinthakumar et al. 2018, Yang, Zhang et al. 2018). Previous studies have highlighted the importance of BFI in low flow analyses, where it has been integrated with hydrograph recession methods to gain a comprehensive understanding of low-flow dynamics (Zhang, Zhang et al. 2017, Sapač, Rusjan et al. 2020, Yang, Li et al. 2020). We used Kling-Gupta Efficiency (KGE) (Gupta, Kling et al. 2009) to evaluate the streamflow prediction accuracy of various Noah-MP-RAPID scenarios and the NWM against observed USGS. The Root Mean Square Error (RMSE) of the bottom 30% of streamflow (Song, Knoben et al. 2024) was also used as a metric to assess the models' performance in capturing the low flow conditions.

2 Materials and Methods

2.1 Study area

We selected the southwestern U.S. due to the low performance of NWM and models in predicting streamflow in this region (Towler et al., 2023). We included 390 USGS gauges across five major USGS two-digit Hydrologic Unit Code (HUC-2) basins (Figure 2): Rio Grande (HUC13), Upper Colorado (HUC14), Lower Colorado (HUC15), Great Basin (HUC16), and California (HUC18). These basins are defined using the NHDPlus version 2 geospatial dataset (Horizon Systems Corporation, 2007), which integrates the National Hydrography Dataset's (NHD) 1:100,000-scale stream network, the 1-arc second National Elevation Dataset (NED), and the Watershed Boundary Dataset (WBD). The NHDPlus version 2 dataset provides NHDPlus catchments, flowlines, and attributes, utilized for generating river connectivity files and computing flowline slopes. According to the NHDPlus v2, the selected region includes 479130 river reaches with an average length and catchment of and 2.29 km and 4.23 km^2, respectively.

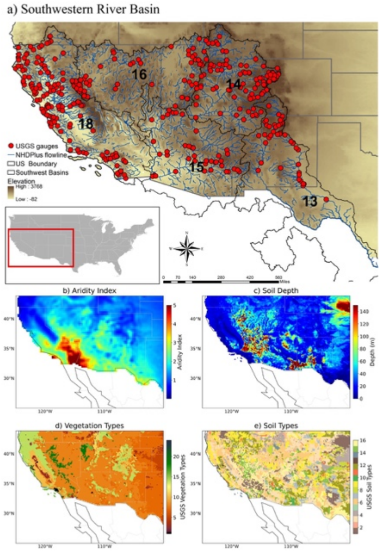

Figure 2. (a) The Southwestern River Basins featuring 390 USGS gauges used in this research, along with Shuttle Radar Topography Mission (SRTM) elevation data; (b) aridity index, indicator of the degree of dryness of a climate at a specific location; (c) soil depth (bedrock); (d) vegetation types; and (e) soil types. For clarity, map (a) shows the NHDPlus flowlines with Strahler order >4 (also in Figures 10)

2.2 Description of Datasets

In this study, we integrated multiple datasets to drive the Noah-MP land surface model and evaluate its performance in simulating hydrological processes. Specifically, we used precipitation data from NLDAS-2, IMERG, and AORC (Section 2.2.1) and atmospheric forcing data from NLDAS-2 (Section 2.2.2) to provide comprehensive meteorological inputs for the model simulations. Soil characteristics were derived from State Soil Geographic (STATSGO) U.S soil map and USGS 24-category vegetation classification datasets. Bedrock depth information was obtained from Shangguan et al. (2017) (Section 2.2.5) to enhance the model's subsurface representations.

2.2.1 Precipitation Data

We used three precipitation datasets to drive Noah-MP scenarios: the North American Land Data Assimilation System (NLDAS-2, (Xia, Mitchell et al. 2012)), the Integrated Multi-satellite Retrievals for GPM (IMERG) Final (Huffman, Bolvin et al. 2020), and the NOAA Analysis of Record for Calibration (AORC,(Fall, Kitzmiller et al. 2023)). The NLDAS-2 precipitation combines observations and model reanalysis to support offline land surface modeling. It includes data from sources like NCEP Stage II/IV analyses and the Climate Prediction Center's gauge-based analysis. NLDAS-2 provides 0.125° × 0.125° spatial resolution data, available hourly and monthly. The IMERG-Final dataset provides half-hourly precipitation at a 0.1° × 0.1° resolution, covering latitudes from 60°S to 60°N (Huffman, Bolvin et al. 2020). It integrates satellite-based passive microwave and infrared sensor data, with regionalization and bias correction applied using GPCC gauge records (Huffman, Bolvin et al. 2020). AORC offers gridded meteorological data across the continental U.S. and Alaska, with a 30-arc-second (~800 meters) resolution and hourly temporal resolution (Fall, Kitzmiller et al. 2023). It compiles weather information for land-surface, snow, and hydrologic models and serves as the forcing data for calibrating the NWM version 2.1 model (Hong, Xuan Do et al. 2022). In this study we use the AORC precipitation while using other atmospheric variables of NLDAS 2, since it used as forcing to derive NWM. Both the IMERG AORC precipitation are mapped to a 0.125° resolution through bilinear interpolation for compatibility with NLDAS-2 (Xia, Mitchell et al. 2012) to match the NLDAS 2 resolution.

2.2.2 Atmospheric Forcing, Soil and Vegetation Parameters

We used the NLDAS-2 forcing dataset to drive the Noah-MP model. This dataset includes hourly data on downward shortwave and longwave radiation, temperature, specific humidity, surface air pressure, and wind speed, with a spatial resolution of 0.125° over the CONUS region. The NLDAS-2 forcing dataset has been extensively used to drive LSMs in our previous modeling studies (Niu, Fang et al. 2020, Agnihotri, Behrangi et al. 2023, Farmani, Behrangi et al. 2024). This study used the State Soil Geographic (STATSGO) U.S soil map and USGS 24-category vegetation classification datasets, aggregated to match the NLDAS-2 resolution, ensuring consistent soil and vegetation parameterization. Noah-MP model's lookup tables (Niu, Fang et al. 2020) assigned the appropriate parameters for soil and vegetation categories.

2.2.3 Streamflow Data for Evaluation and Comparison

To evaluate the model's performance in modeling BaseFlow Index (BFI) and predicting streamflow, we used observed daily streamflow data from the US Geological Survey (USGS) for 390 gauges across the southwestern United States (Figure 2). The selected gages cover wide ranges of drainage area, varying from 10.51 km2 (Kings Cyn CK NR Carson City, NV - 10311100) to 638432.06 km2 (Colorado R Blw Yuma Main Canal WW at Yuma, AZ - 09521100)

We also used NWM (v2.1) retrospective (https://noaa-nwm-retrospective-2-1-pds.s3.amazonaws.com/index.html) daily streamflow data for comparison and to evaluate potential improvements. NWM integrates the WRF-Hydro model to simulate the water cycle and deliver streamflow predictions for 2.7 million reaches across the contiguous United States (Salas, Somos-Valenzuela et al. 2017, Cosgrove, Gochis et al. 2024). It combines numerical weather forecasting models with Noah-MP to generate runoff on a 1-km grid, employing Muskingum–Cunge channel routing techniques to predict streamflow for NHDPlus stream reaches (Shastry, Egbert et al. 2019).

2.2.4 SoilGrids250m Dataset

We used the SoilGrids250m dataset to derive the saturated hydraulic conductivity and soil water retention curve parameters using the Pedo-Transfer function (PTF) proposed by Wösten et al. (1999) (see Appendix A). The SoilGrids250m product, developed by ISRIC (Poggio, de Sousa et al. 2021), represents a significant advancement over the previous machine learning-generated SoilGrids1km dataset (Hengl, de Jesus et al. 2014). It provides global estimates of various soil characteristics, including the percentages of clay, sand, and silt, as well as organic carbon content and soil bulk density. Compared to the SoilGrids1km dataset, SoilGrids250m provides improved accuracy for soil texture and other characteristics (Oloruntoba, Kollet et al. 2024).

2.2.5 Bedrock Depth

This study utilizes the dataset produced by Shangguan et al. (2017) which presents a comprehensive framework for estimating the global depth to bedrock (DTB). This dataset integrates observations from approximately 1.3 million borehole records and 130,000 soil profile locations enhanced with pseudo-observations to improve global coverage (Shangguan, Hengl et al. 2017). Leveraging a diverse array of covariates, including DEM-based hydrological indices, lithologic maps, and MODIS satellite products, the authors employed machine learning techniques such as Random Forest and Gradient Boosting to predict DTB at a fine resolution of 250 meters.

2.2.6 Noah-MP with Advanced Soil Hydrology

In this research, Noah-Multiparameterization Land Surface Model (Noah-MP LSM) (Niu, Yang et al. 2011) was chosen for its widespread use in the Weather Research and Forecasting (WRF) model (Skamarock, Klemp et al. 2008), the Unified Forecast System (UFS) (Moon, Knutson et al. 2022), and NWM. These models are vital for weather, short-term climate projections, and streamflow forecasting. Noah-MP distinguishes between bare and vegetated areas, ensuring accurate calculations of surface energy and fluxes (Agnihotri, Behrangi et al. 2023).

The Noah-MP version employed in this research incorporates several significant features and advancements to enhance the simulation of hydrological and ecological processes. These include advanced plant hydraulics, improved soil hydrology with explicit representation of soil water movement, the explicit prediction of plant water storage, integration of a mixed-form Richards' equation for simulating surface ponding and preferential flow, and the consideration of variability in infiltration capacity due to soil macropores.. It explicitly predicts plant water storage, which is computed as the residual of root water uptake driven by the hydraulic gradient between the soil and roots and transpiration (Niu, Fang et al. 2020). Furthermore, it integrates a mixed-form Richards' equation that simulates surface ponding, infiltration, and preferential flow (Niu, Fang et al. 2024). The soil hydrology simulates the movement of water from the bedrock to the vegetation canopy to fulfill plant transpiration requirements. It additionally considers the variability of infiltration capacity by considering fractional area of preferential flow pathways created by soil macropores in the fields. Table 1 details consistent optional schemes across experiments, including surface layer exchange, radiation transfer, phase changes, and runoff scheme.

2.2.6.1 Optional Soil Hydraulics Schemes

The version of Noah-MP used in the current study offers optional hydraulic models using Van Genuchten-Mualem (VGM) and Brooks-Corey with Clapp-Hornberger (BC/CH) parameters. Generally, the Van Genuchten water retention curve produces less suction than BC/CH in the drier end of soil moisture (see

Niu et al., 2020) A polynomial function is applied to smooth the BC/CH water retention curve for better convergence near saturation (Bisht, Riley et al. 2018)

2.2.6.2 Representing Preferential Flow

The enhanced Noah-MP incorporates a dual-permeability model (DPM) to represent preferential flow, partitioning the grid into macropore and matrix flow domains. This approach is based on the work of Simunek and van Genuchten (2008) and Gerke and van Genuchten (1993a, 1996). This model accounts for subgrid variability in infiltration capacity and water transfer between the two subgrid domains, including “lateral infiltration” and lateral movement of surface ponded water. The overall water content ($\theta\ [\frac{m^3}{m^3}]$) and vertical water flux ($q\ [\frac{m}{s}]$) for a grid cell are calculated using the expressions $\theta = F_a\ \theta_a + (1-F_a)\ \theta_i$ and $q = F_a\ q_a + (1-F_a)\ q_i$. Here, F denote the fraction of soil grid and the subscripts a and i refer to macropore and micropore domains, respectively. This method also applies to other water fluxes, such as soil evaporation (E_{soil}) and groundwater recharge.

2.2.6.3 The Mixed-Form Richards’ Equation

Most LSMs utilize a mass-based (θ-based) Richards' Equation (RE) for unsaturated soils, as noted by Chen & Dudhia (2001) and Oleson et al. (2010). However, this approach often struggles to accurately describes saturated conditions such as surface ponding and groundwater dynamics. In contrast, the current Noah-MP incorporates the method developed by Celia et al. (1990), which involves solving the mass-pressure (θ-h) mixed-form RE. This solver can compute the pressure head (h, m), continuously across saturated and unsaturated zones, while conserving mass (θ) and employs an adaptive time-stepping scheme to enhance accuracy. All the model experiments in this study benefit from using Mixed-Form Richard’s equation solver (Niu, Fang et al. 2024).

2.2.6.4 Surface Ponding

Surface ponding arises when the pressure head of the surface layer exceeds the air entry pressure. As a result, the upper boundary condition (BC) transitions from flux BC to head BC. Infiltration-excess runoff occurs when the depth of surface ponding(H_{top}, mm), surpasses a specified threshold ($H_{top,max}$, mm). This situation leads to the connection and runoff of water at local depressions within a grid cell. The model's vertical domain extends to the depth of the bedrock, with a lower boundary condition of zero-flux. Groundwater discharge is simulated using the TOPMODEL concept, which is based on the water table depth determined by interpolating the model predicted pressure head at specified layers.

Table 1 Noah-MP Options used in this study.

Process	Options	Schemes
Dynamic vegetation	DVEG = 2	Dynamic vegetation
Canopy stomatal resistance	OPT_CRS = 1	Ball-Berry type
Moisture factor for stomatal resistance	OPT_BTR = 1	Plant water stress
Runoff and groundwater	OPT_RUN = 1	TOPMODEL with groundwater
Surface layer exchange coefficient	OPT_SFC = 1	Monin-Obukhov similarity theory (MOST)
Radiation transfer	OPT_RAD = 1	Modified two-stream
Ground snow surface albedo	OPT_ALB = 3	Two-stream radiation scheme (Wang, He et al. 2022)
Precipitation partitioning	OPT_SNF = 5	Wet bulb temperature (Wang, Broxton et al. 2019)

Lower boundary condition for soil temperature	OPT_TBOT = 2	2-m air temperature climatology at 8m
Snow/soil temperature time scheme	OPT_STC = 1	Semi-implicit
Surface evaporation resistance	OPT_RSF = 1	Sakaguchi and Zeng (2009)
Root profile	OPT_ROOT = 1	Dynamic root (Niu, Fang et al. 2020)

2.2.7 Model Experiments

2.2.7.1 General attributes

We conducted 10 experiments using an updated Noah-MP model, focusing on three categories related to baseflow uncertainties in 1) representations of hydrological processes, 2) soil water retention curve parameters (hydraulic parameters), and 3) precipitation datasets. All 10 experiments were driven by the same atmospheric variables from NLDAS-2 forcing data, downward shortwave and longwave radiation, temperature, specific humidity, surface air pressure, and wind speed, at 0.125° resolution, with initial conditions from model spin-up runs from 1980 – 2019 for three iterations. The first two iterations starting from soil moisture at 0.3 m³/m³ and soil temperature at 287K served as model spin-up (80 years). All the experiments in the hydrological processes and hydraulic parameters scenarios were driven by the NLDAS-2 precipitation data. For the precipitation datasets scenarios, the experiments were driven by the by the three precipitation datasets: IMERG, AORC, and NLDAS-2, all with other atmospheric variables from NLDAS-2, with 11 iterations (10 as model spin-up) from 2014 – 2019.

Scenario parameters followed Niu et al. (2020), with adjustments for the Moderate Resolution Imaging Spectroradiometer (MODIS) leaf area index data. No calibration was done for dual-domain schemes related to preferential flow and ponding depth (Šimůnek and van Genuchten 2008). Experiments used a uniform soil layer thickness setup with varying number of vertical layers (5 – 15 layers), depending on the bedrock depth with a maximum depth of 49.0 meters in this region (Pelletier et al., 2016), and a minimum depth of 4.0 meters.

Table 2. Model Experiments configurations

Category	Experiment name	Soil Moisture Solver	Ponding depth (mm)	Soil Hydraulics	Forcing	Soil Water Retention Characteristics Parameters
Hydrological Process	CH	Mixed Form RE	50	Brooks-Corey/Clapp-Hornberger	NLDAS-2	Noah-MP Table
	VGM	Mixed Form RE	50	Van-Genuchten		
	VGM0	Mixed Form RE	0	Van-Genuchten		
	DPM	Dual Permeability, Mixed Form RE	50	Van-Genuchten		
Hydraulic Parameters	ML	Mixed Form RE	50	Van-Genuchten	NLDAS-2	ML-Based (Gupta et al., 2022)
	PTF50	Mixed Form RE	50			PFT (Wösten et al., 1999)
	DPMPTF0	Dual Permeability, Mixed Form RE	0			PFT (Wösten et al., 1999)
Precipitation	NLDAS	Mixed Form RE	50	Van-Genuchten	NLDAS-2	Noah-MP Table
	IMERG				NLDAS-2, IMERG	
	AORC				NLDAS-2, AORC	

2.2.7.2 Experiments Developed in this Study

The hydrological processes scenario consists of four experiments to evaluate the impact of different representations of hydrological processes (or models) on baseflow generation and BFI (Table 2). These models vary by soil hydraulic schemes (Brooks-Corey with Clapp-Hornberger parameters and Van-Genuchten), ponding depth threshold (50 mm and zero), and model physics (single vs. dual permeability) (see Niu et al., 2024 for details). To assess the impacts of soil hydraulic properties on BFI, we conducted two experiments using a 50 mm ponding threshold with a single domain scheme: one with the Brooks-Corey and Clapp-Hornberger (CH) parameters, whereas the other with Van-Genuchten (VGM). This comparison allows us to isolate the effect of different soil retention characteristics on baseflow generation.

A third experiment incorporated the Dual-Permeability Model (DPM) within the VGM framework, maintaining the 50 mm ponding threshold (referred to as DPM). This setup examines the effects of macropore flow and preferential pathways on baseflow, helping us understand how soil structure influences hydrological responses. The fourth experiment explored surface ponding's role in baseflow generation by setting the ponding threshold ($H_{top,max}$) to 0 mm in a VGM framework (VGM0). By eliminating the potential for surface ponding, we aimed to assess its impact on infiltration and subsequent baseflow processes.

It should be noted that we first evaluated the BFI from hydrological processes scenarios. based on their performance we selected the scenario which has the closest BFI pattern to USGS as basic configuration for hydraulic parameters and precipitation forcing experiments. Also, macropore volume fraction was determined using modeled soil organic matter (SOM) from Noah-MP with a microbial-enzyme model (Zhang, Niu et al. 2014). The macropore volume fraction ranged from 0.05 to 0.15.

For hydraulic parameters scenarios, we used the VGM configuration with two datasets instead of the Noah-MP lookup table. The first dataset used machine learning-generated parameters (ML scenario) by Gupta et al. (2022), and the second used parameters derived from the Pedo-Transfer Function (PTF) and SoilGrids250 data (PTF50 scenario). The third scenario employed the DPM framework with a zero ponding threshold and PTF-generated soil water retention curve parameters (Table 2). To evaluate the effect of precipitation forcing datasets, three VGM configuration experiments were conducted with AORC, IMERG, and NLDAS-2 datasets (Table 2), covering 2014-2019 due to limitation in the length of IMERG precipitation and our storage space.

2.2.8 RAPID Routing Model

The Routing Application for Parallel computation of Discharge (RAPID, David, Maidment et al. (2011)) used a matrix formulation of the Muskingum method to calculate discharge across river networks with numerous river reaches. The National Oceanic and Atmospheric Administration (NOAA) National Water Model incorporated RAPID as an alternative river routing model, and it served as a component of the Streamflow Prediction tool (Snow, Christensen et al. 2016).
For this study, we used daily-mean gridded lateral inflow, specifically surface and subsurface runoff outputs from the final loop of the 1980–2019 Noah-MP simulations, as input for the RAPID model to simulate daily discharge within the river network. We developed a consistent vector-based river network for the study region (Figure 2a) using the NHDPlus V2 dataset, which was previously applied to create the river network for the entire Mississippi River Basin (MRB) (Tavakoly, Snow et al. 2017, Tavakoly, Gutenson et al. 2021). To ensure consistency in the streamflow results, we also applied the same parameters across all RAPID simulations.

2.3 BFI Calculation

Streamflow consists of two main components: quick flow and baseflow. Quick flow, also known as storm flow, results from faster streamflow pathways such as direct precipitation and surface runoff due to infiltration excess and saturation overland flow (Hall 1968, van Dijk 2010). In contrast, baseflow is derived from groundwater, which is the same as subsurface runoff in Noah-MP, and delayed sources like snowmelt, ensuring water availability during dry periods (Hall 1968, van Dijk 2010).

While chemical tracers can estimate baseflow (Genereux 1998), they are labor-intensive and costly (Xie, Liu et al. 2020). Therefore, non-tracer methods like graphic methods (Sloto and Crouse 1996, Arnold and

Allen 1999) and digital filtering techniques (Chapman 1991, Furey and Gupta 2001, Huyck, Pauwels et al. 2005, Tularam and Ilahee 2008, van Dijk 2010) have been developed to estimate baseflow from streamflow data without extensive fieldwork (Kissel and Schmalz 2020). Many studies have compared these methods, showing comparable performance (de Roo, Beck et al. 2015, Kissel and Schmalz 2020, Chen and Ruan 2023).

Selection of methods for calculating baseflow has a significant impact on the resulting values. Given that the ground truth values of BFI are unknown, the method selection process involves a degree of subjectivity (Beck, van Dijk et al. 2013, de Roo, Beck et al. 2015). Comparative studies across various catchments often reveal strong correlations between different techniques for determining BFI (Chapman 1999, Eckhardt 2008, de Roo, Beck et al. 2015, Kissel and Schmalz 2020). Eckhardt (2008) calculated BFI values for 65 catchments using seven different methods and observed coefficients of determination (R^2) of 0.85 or higher, indicating strong agreement among the techniques. Regardless of the approach used for baseflow separation and BFI computation, the results tend to be highly correlated and consistent (Eckhardt 2008, Beck, van Dijk et al. 2013, de Roo, Beck et al. 2015).

In this study we used the Van Dijk (2010) baseflow separation method, based on a linear reservoir model, to compute baseflow and BFI for each catchment. According to Van Dijk (2010), the linear reservoir model performs as well as the two-parameter model but with the advantage of using a single parameter, thereby reducing parameter variability. Chapman (1999) also supports the use of this model, highlighting its efficiency in baseflow estimation. We aim to deepen our understanding of baseflow generation by applying a consistent separation method across both benchmark data and various scenarios, ensuring uniformity in our analysis. Hence, the uncertainty of selecting the separation method does not affect the purpose of this study, which uses BFI to understand and improve baseflow generation processes. However, the Van Dijk (2010) separation model which describes falling limb of hydrograph as an exponential decay factor of time, could capture the exponential factor in TOPMODEL with groundwater option in Noah-MP. As described in Van Dijk (2010) and Beck et al (2013), single BFI and k values were calculated from the Q record for each catchment. For linear reservoir model, relation between the streamflow and reservoir storage can described as:

$$Q(t) = kS(t) \tag{1}$$

Where Q (mm d^{-1}) is the streamflow, k (d^{-1}) is recession coefficient and S (mm) is reservoir storage. Also, from the continuity equation we have:

$$\frac{dS}{dt} = -Q(t) \tag{2}$$

Let’s take a derivation from equation (1):

$$\frac{dQ(t)}{dt} = k\frac{S(t)}{dt} \tag{3}$$

Now replace equation (3) in equation (2) and solve it:

$$\frac{dQ(t)}{k\,dt} = -Q(t) \tag{4}$$

$$Q(t) = Q\ (t-1)\ \exp{(-k)} \tag{5}$$

Equation (5) describes the falling limb of hydrographs, and the k for a specific catchment was calculated as follows:

$$k = -\ln{(\frac{Q(t)}{Q_*})} \tag{6}$$

The parameter k can be estimated by fitting equation (6) to the data pairs of $Q(t)$ (mm d^{-1}) and Q_* (mm d^{-1}), derived from $Q(t)$ and $Q(t-1)$, as outlined by van Dijk (2010) and Beck et al(2013).

$$Q = \exp\left(\overline{\ln\left(Q(t = 1,2,..,N)\right)}\right) \tag{7}$$

$$Q_* = \exp\left(\overline{\ln\left(Q(t = 0,1,..,N-1)\right)}\right) \tag{8}$$

Where N denotes the Nth observation in data pairs. To create the data pairs, we first removed zero values and days showing an increase from the previous day. Additionally, the five days following these instances were excluded to account for quick flow (Beck, van Dijk et al., 2013; de Roo, Beck et al., 2015). The multi-start downhill simplex algorithm was employed, with Root Mean Square Error (RMSE) as the fitting criterion, to find the optimum fitting parameter k.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}\left(Q_{pre} - Q\right)^2}{N}} \tag{9}$$

where Q_{pre} is predicted streamflow using Eq. (5). The k value obtained from the calculations was utilized to separate the Q record into baseflow and quick flow, employing a combination of forward- and backward-recursive digital filters (van Dijk 2010) for each day. Finally, the ratio of total baseflow to total streamflow estimates the BFI. To validate our implementation of the baseflow separation method proposed by Van Dijk (2010), we applied the method to a station from the Global Runoff Data Center (GRDC) and compared our results (Figure S1) with those obtained by Beck et al. (2013), who also implemented the method. Our estimation of baseflow and BFI agrees closely with Beck et al. (2013).

3 Results

3.1 BFI

3.1.1 The Effects of Representations of Hydrological Processes

The BFI values derived from the 40-year streamflow data at the USGS gauges displays a wide range from 0 to 0.95 with a median of 0.78 (Figures 3a and 4a), consistent with the those of Beck et al. (2013) and Beck et al. (2015). Lower BFI values appear in the drier southwestern coastal and southern regions of the domain, whereas higher values are in the wetter northwestern Rocky Mountains. The study domain shows a gradient from the higher BFIs in the northern wetter climates to lower BFI values in the southern dry climates, reflecting the effects of the varying hydroclimate conditions on soil formation processes through weathering and sediment transport and deposition (Figure 2). The BFI pattern over our study region was close to the aridity index pattern (Figure 2b), where areas with a higher aridity index displayed lower BFI, and those with a lower index had higher BFI. Additionally, the observed BFI gradient across the study domain may have been related to variations in bedrock depth (Figure 2c). In the Rocky Mountains, the presence of shallower bedrock likely contributed to thinner soils, which could facilitate baseflow generation. Furthermore, the steeper terrain in these regions promoted faster water movement through the soil, further enhancing baseflow generation (Figure 2a).

NWM overestimated BFI across most gauges, including both low- and high-BFI regions, with a median value of 0.88 (Figure 3b). The BFI boxplot of the NWM indicates that more gauges exhibit higher BFIs and a narrower range of values compared to the USGS (Figure 4a). The CH experiment, with a median BFI of 0.82, also overestimated BFI, more obviously in areas with low BFIs (Figure 3c and 4a). Similar to NWM's distribution, CH also showed a narrower range when compared with that of the USGS data (Figure 4b). Additionally, CH generated lower BFI values than the USGS in the high baseflow regions. NWM uses the Brooks-Corey hydraulics as did the CH experiment but with optimized hydraulic parameters.

VGM produced a median BFI of 0.76 (Figure 4a) and a wider BFI range than CH (Figure 4), closer to the distribution of the USGS, especially in the southern low BFI regions (Figure 3). VGM performed better than CH in low BFI areas (Figure 3), and so did BFI range and distribution from VGM compared with to those from the USGS (Figures 4a and 4b). However, the VGM slightly underestimated BFI in high-BFI regions over the Rockies (Figure 4b), likely because Noah-MP did not account for lateral flow caused by the steep topography in this region. By excluding lateral flow, the model underrepresented water redistribution across slopes, which reduced infiltration and limited groundwater recharge.

Compared with VGM, DPM produced a median BFI of 0.81 and a much narrower range of BFI distribution (Figure 3e). It significantly overestimated BFI in the southern low-BFI areas but captured the high BFI values over the Rockies slightly better (Figures 4a and 4b). The spatial distribution of organic matter may be not representative enough, because both VGM and DPM used the VGM hydraulic parameters and the same ponding depth, and the only difference was the macropore volume fraction represented by soil organic matter in DPM.

The inclusion of ponding process significantly impacted baseflow generation. VGM0 produced a lower median BFI of 0.65 compared to VGM with a 50 mm ponding depth threshold (Figures 3d and 3f). Although VGM0 produced a similar range of BFI compared to the USGS, it significantly underestimated high BFI, shown by the low density of BFI values above 0.7 (Figure 4b). This indicates that a model without consideration of surface ponding (allowing all infiltration-excessive water to run off) would underestimate BFI, and a spatially constant threshold is not representative (should include spatially-variable microtopography information).

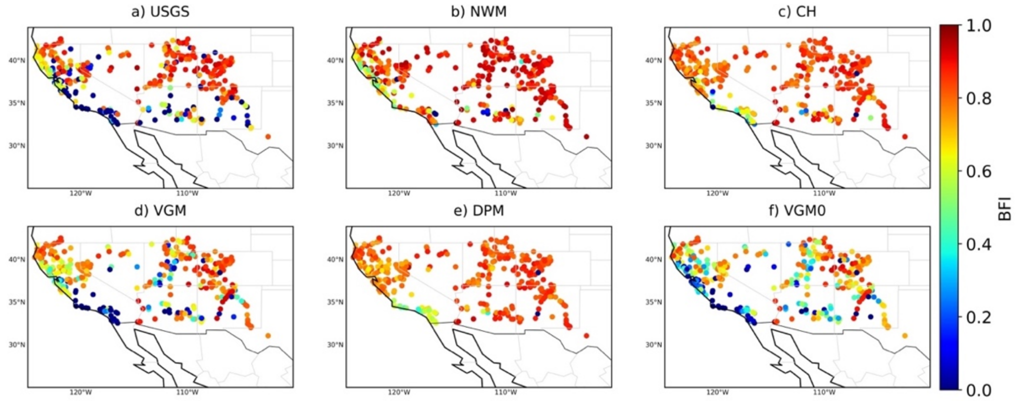

Figure 3. BFI at 390 gauges across the southwestern region (a) derived from the USGS streamflow as well as those from the model experiments (b) NWM; (c) CH; (d) VGM; (e) DPM; and (f) VGM0.

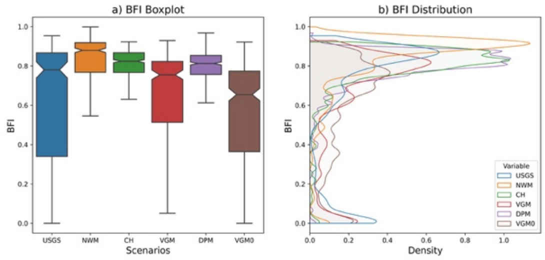

Figure 4(a) Boxplot (b) and Distribution of BFI for USGS, NWM, and different scenarios.

3.1.2 The Effects of Hydraulic Parameters

To analyze the impact of selecting parameters for soil water retention curve, we used VGM, which matches the closest the USGS benchmark. We applied the VGM configuration with different datasets for the hydraulic parameters scenario.

All experiments with hydraulic parameters generally captured the pattern seen in the USGS data. They showed low BFI values in the southwestern coastal regions and high BFI values in the northeast Rockies (Figures 5). Using the ML-derived parameter dataset increased BFI, with a median of 0.78, especially in the southwestern coastal areas compared to VGM with the parameters from the look-up table. BFI also increased at most gauges in the Colorado State, where the BFI values from VGM are lower than the USGS data (Figures 5a, and b). However, the ML-based parameters degraded the simulation in the southern California and Arizona compared to VGM. Overall, the ML-based parameters result in a narrower range of BFI distribution by increasing BFI at most gauges (Figure 6a), with the reduced density of BFI below 0.1 (Figure 6b) and increased density of BFI values above 0.7 compared to VGM. VGM performed better than ML in capturing the higher BFI in the northeast (Figure 6b). The difference between ML and VGM could be related to the generally larger saturated hydraulic conductivity (Figure S2).

PFT50, soil water retention curve parameters derived from PFT (Wösten, Lilly et al. 1999), produced a median BFI of 0.78 with a pattern similar to that of ML, but with a smaller increase in California. PTF50 also increased BFI, particularly in the southwestern regions, compared to VGM. This is similar to ML but with slightly better performance and a wider range of BFI (Figure 6a). Like ML, PTF50 increases BFI in central regions (Figures 5a, and b). However, VGM again performed better than PTF50 in capturing the highest BFI in the northeast regions (Figure 6b).

PTFDPM0 uses the dual permeability model with the PTF-derived parameters but with the ponding threshold equals zero to produce model surface runoff (Table 2). PTFDPM0, with a median BFI of 0.74, shows a pattern similar to VGM, PTF50, and ML, predicting slightly higher BFI values for high BFI regions (Figure 5d). It slightly captures the high BFI from the USGS gauges (Figure 6b), thereby a little enhancing the VGM's ability to predict the highest BFI, though it still underestimates the high BFI regions. Overall, using different hydraulic parameters does not significantly affect the BFI, although it slightly shifts the BFI toward higher values. However, the hydraulic parameters may influence the timing and peak value of the streamflow, which will be analyzed in Section 3.2.

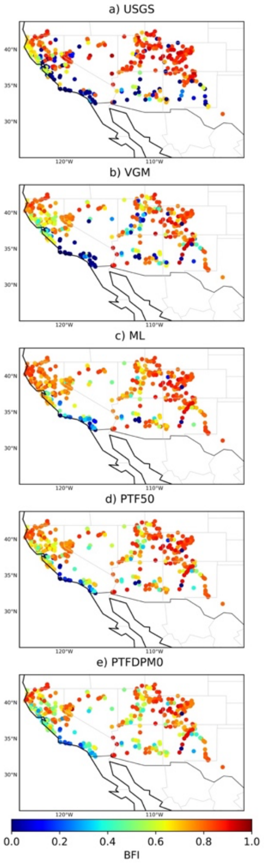

Figure 5 BFI at 390 locations across the southwestern region for the hydraulic parameters scenarios. Panels show the following: (a) USGS; (b) VGM; (c) ML; (d) PTF50; and (e) PTFDPM0.

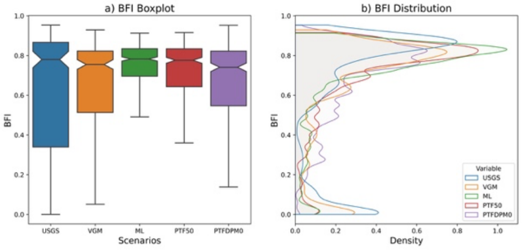

Figure 6 Boxplot (a) and Distribution of BFI for USGS, VGM, and hydraulic parameter scenarios.

3.1.3 The Effects of Precipitation Forcing

Similar to the hydraulic parameter scenarios, we used the VGM model with three precipitation products. Due to limitations in the length of IMERG precipitation, these experiments were run for the period 2014 – 2019.

IMERG, using the IMERG precipitation, produces a BFI median of 0.74, with lower BFIs at the stations in the coastal regions of the middle California River Basin, reducing the high biases by NLDAS (Figures 3a, 7a, c). IMERG effectively captured the low BFIs that NLDAS and VGM fail. Also, IMERG produced a density closer to USGS for BFI values lower than 0.2. Additionally, IMERG increases BFIs for some northeastern stations with high BFIs (red square in Figure 8b), resulting in a wider range (Figure 8a).

AORC, using the AORC precipitation, also shows a BFI spatial pattern similar to NLDAS (Figure 1d). However, it predicts higher BFI values for some stations, leading to a higher BFI median of 0.77 compared to NLDAS and IMERG. This is reflected in both the boxplot and distribution plot (Figure 2a, b), where AORC had the highest BFI median among the scenarios. The BFI distribution for AORC (Figure 8b) showed a narrower range. Among all the forcing scenarios, AORC performed the worst due to the overestimated BFI in low-BFI regions. In contrast, IMERG performed the best by accurately capturing the low BFI regions along the coast, where NLDAS and VGM have neglected.

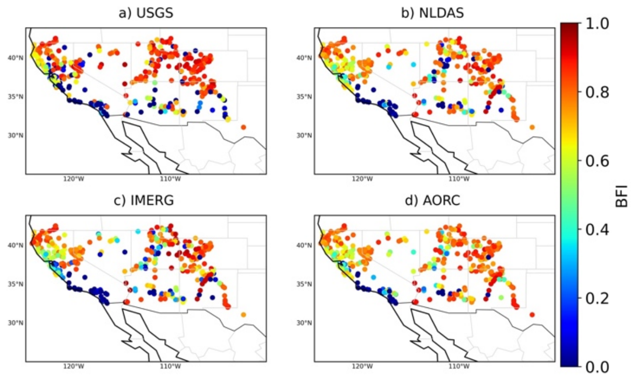

Figure 7. BFI at 390 locations across the southwestern region for the forcing scenarios. Panels show the following: (a) USGS; (b) NLDAS; (c) IMERG; and (d) AORC. The BFI for USGS is computed for 2014-2019 to be consistent with forcing scenarios.

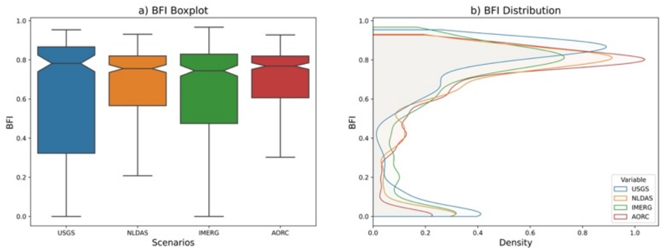

Figure 8.BFI Boxplot (a) and (b) distribution of BFIs for USGS, and, and forcing scenarios. The BFI for USGS and VGM are computed for 2014-2019 to be consistent with forcing scenarios.

3.2 Evaluation of Streamflow Prediction Skill with KGE and Low Flow RMSE

We also evaluated the model performance in simulating streamflow using the median KGE and low flow RMSE for the various model configurations. Among all the physical process and hydraulic parameters experiments, ML showed the best performance with the highest median KGE of 0.29, the lowest low flow RMSE of 1.57, and the most stations of positive KGE values of 272 (Table 3). VGM and PTF50 also performed well, each with a median KGE of 0.28 and a low flow RMSE of 1.62, indicating their reliability in simulating hydrological processes. ML also has the smallest KGE range (Figure 9a), with its distribution skewed toward higher KGE values (Figure 9b), making it a robust option across different regions.

The PTFDPM0 scenario exhibited the poorest performance with the lowest median KGE of 0.06, the highest low flow RMSE of 2.62, and the fewest stations with positive KGE, despite occasionally capturing the highest KGE among all scenarios (Figure 9b). This suggests that while PTFDPM0 can perform well in some instances, its overall reliability is significantly lower than other configurations.

The most negative KGE stations are located in the Great Basin, Upper Colorado, and Rio Grande HUC2 regions (Figure S3), which correspond to high BFI regions. In contrast, the most positive KGE stations are found in the California River Basin (Figure S3). Notably, some stations in the Rio Grande and Upper Colorado regions have negative KGE values resulting from NWM but positive KGE from ML (Figure S3a, b). This further highlights the improved performance of the ML configuration, especially in regions where other configurations struggle.

Table 3 Median KGE, Number of stations with positive KGE, and low flow RMSE for scenarios covering 1980–2019

Scenario	Median KGE	Number of stations with positive KGE	Low flow RMSE
NWM	0.16	221	2.35
CH	0.17	227	2.07
VGM	0.28	257	1.62
DPM	0.21	229	1.77
VGM0	0.13	211	2.50
ML	**0.29**	**272**	**1.57**
PTF50	0.28	257	1.62
PTFDPM0	0.06	200	2.62

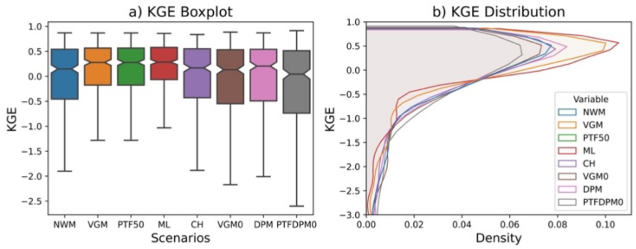

Figure 9. Boxplot (a) and Distribution of KGE for NWM, and physical process and hydraulic parameter scenarios.

Regarding the precipitation forcing scenarios, IMERG performed the best with the highest median KGE (0.30), the largest number of positive KGE stations (309), and the lowest low flow RMSE (0.99), though over a shorter length of period (Table 4). Or at least, using IMERG precipitation data can enhance KGE and low flow predictions. Integrating IMERG data with the ML configuration, ML_IMERG, could further improve baseflow predictions, combining the strengths of both approaches for more accurate hydrological modeling (Figure S4, S5).

Table 4 Median KGE, Number of stations with positive KGE, and low flow RMSE for the scenario covering 2014–2019

Scenario	Median KGE	Number of stations with positive KGE	Low flow RMSE
AORC	0.13	221	1.63
IMERG	**0.30**	**309**	**0.99**
NLDAS	0.19	262	1.13

We also computed KGE improvement (%) to evaluate the enhancement of different configurations upon NWM for 1980 – 2019 (Table 5). For each gauge station, we calculate the improvement in KGE as a percentage increase relative to the NWM. This is computed using the formula:

$KGE_{Imp} = \frac{KGE_{i,j} - KGE_{i,NWM}}{KGE_{i,NWM}}$ for station i under scenario j.

Among all scenarios, ML shows the highest improvement, with a median of 20%. Notably, 294 stations shows improvements under the ML configuration (Figure 10), making it the most effective in improving model performance across a wide range of stations, which are scattered throughout the entire region with a wide range of soils, vegetation, and climates. There are 79 stations showing an improvement more than 100% over NWM in the southwestern areas, which are characterized by low BFIs. The use of the VG scheme with ML parameters not only improves low flow predictions in Southern California coast and Arizona but also significantly enhances KGE at these locations.

VGM and PTF50 scenarios also outperform NWM, each achieving a 15% improvement in median KGE. Both scenarios show KGE improvements at 282 stations, indicating their robustness in enhancing model accuracy. These improvements are widely distributed, suggesting that these configurations can be reliable alternatives to NWM for various hydroclimate conditions.

On the other hand, CH, DPM, VGM0, and PTFDPM0 show negative median KGE improvements, with more stations experiencing degraded KGE values than improvements. For example, the VGM0 scenario had a -5% median KGE improvement, with only 91 stations showing improvement, while 299 stations had degraded KGE values. Similarly, PTFDPM0 showed a -2% median KGE improvement, with 97 stations improving and 293 degrading, reflecting its overall weaker performance. While ML stands out as the best performing scenario with significant KGE improvements across the region, VGM and PTF50 also provide substantial enhancements over the NWM scenario. However, scenarios like VGM0 and PTFDPM0 indicate the importance of parameter selection and configuration in achieving reliable model performance, as they show more stations with degraded KGE values than improvements.

Furthermore, the map in Figure 10 illustrated the spatial distribution of KGE improvements, with green dots representing stations where KGE improved and red dots where KGE degraded. The high concentration of green dots underscored the effectiveness of VGM scheme with ML parameters, especially in regions where traditional models like NWM struggled.

Table 5 Median KGE improvement against NWM and number of stations with improved performance for the scenarios covering 1980–2019.

Scenario	Median KGE improvement (%)	Number of stations with improved KGE	Number of stations with degraded KGE
CH	-9	160	230
VGM	15	282	108
DPM	-4	173	217
VGM0	-5	91	299
ML	**20**	**294**	**96**
PTF50	15	282	108
PTFDPM0	-2	97	293

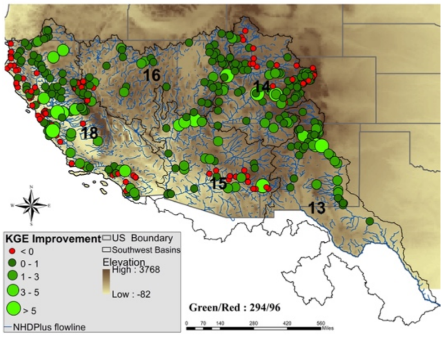

Figure 10 KGE Improvement of ML against NWM. ML was chosen since it outperformed other hydrological process and hydraulic parameter scenarios.

4 Discussion

Modeling the baseflow generation is challenging, because it essentially involves almost all hydrological processes including surface infiltration, surface runoff, percolation, groundwater recharge and discharge.

This study aims to examine how baseflow generation processes in the LSMs affect the accuracy in streamflow predictions, especially in dry regions. Given the importance of baseflow as the primary source of streamflow in these areas, we hypothesize that the baseflow generation processes in the large-scale hydrological models contribute to the inaccuracies in streamflow predictions. By addressing the baseflow issue directly, we indirectly address the groundwater recharge and discharge problem in LSMs. To address the baseflow generation processes, we used a version of Noah-MP that is more physically-based, incorporating the mixed-form Richards equation and a dual domain model, and conducted a series of model experiments with varying soil hydraulic schemes, soil water retention curve parameters, and precipitation datasets. The results in this study provided important insights into the hydrological dynamic of the region and the effectiveness of different modeling approaches.

4.1 Impacts of Hydrological Processes on Baseflow Generation Mechanism

4.1.1 Soil Water Retention Scheme (CH and VGM Experiments)

The accuracy of continental-scale hydrological models in predicting streamflow, particularly in dry regions, hinges on their ability to simulate various water balance components, including surface and subsurface runoff, soil moisture retention, as well as groundwater recharge and discharge. Runoff, which includes both surface and subsurface components, plays a critical role in the distribution of water between immediate streamflow and groundwater recharge, impacting baseflow generation. In dry regions, where precipitation is sparse, subsurface runoff becomes especially relevant, as it often contributes to baseflow—a key element in sustaining streamflow during dry periods. Therefore, examining both surface and subsurface runoff provides insight into how different physical process representations in models affect baseflow estimates.

Comparing the recharge and subsurface runoff for CH and VG models revealed that the CH model exhibits higher recharge and greater subsurface runoff than the VG model. Since subsurface runoff is a primary source of baseflow, the higher values in the CH model led to an overestimation of baseflow (Figures S6 and S7). Moreover, the VG model showed higher surface runoff than the CH model, meaning less water contributes to baseflow (Figure S8). Therefore, the CH model's tendency to overestimate baseflow and the Baseflow Index (BFI) can be attributed to its higher simulated recharge and subsurface runoff.

In low BFI regions, the higher surface runoff produced by VG results in less water infiltrating into the deep soil and recharging the groundwater (Figures S6 and S8). Hence, VG could generate lower baseflow and BFI in these regions (Figure 3d). In contrast, CH shows more recharge into groundwater and consequently higher BFI (Figures S6 and 3c).

4.1.2 Soil Macropores and Ponding Depth (DPM, VGM, and VGM0 Experiments)

The presence of soil macropores in DPM experiment facilitated rapid infiltration and preferential flow through the unsaturated zone, groundwater recharge, and thus baseflow (Mohammed, Cey et al. 2021). In our study, DPM tends to overestimate BFI in low-BFI regions whereas better captured high BFI values. The presence of macropores increased drainage from the surface to the root zone, potentially reducing surface layer moisture retention and increasing groundwater recharge and discharge. This enhances predictions in wet, high-BFI regions but may produce unrealistic results in low-BFI areas. Hence calibration of the soil macropore volume fraction, which is parameterized as a linear function of soil organic matter, is critical to achieve more realistic results. Also, the relationship between soil macropore volume fraction and soil organic matter and other coarse materials (gravels, stones) are worth further investigation.

Including a ponding threshold, as seen in the VGM scenario with a 50 mm maximum ponding depth, is crucial for improving baseflow generation, especially in high-BFI regions. This configuration allows more water to remain on the surface longer for infiltration before running off, which enhances infiltration and groundwater recharge and positively impacts baseflow generation (Figure 4d, f).

A spatially variable ponding threshold should be developed based on the microtopography information. To simulate surface water in floodplains, a parameter, maximum ponding height, is needed to determine when infiltration-excess runoff from surface ponding occurs. To derive this parameter, we will use the standard deviation of subgrid topography for each grid cell, representing microtopographic features, to fit a non-linear relationship. For evaluation, we will compare floodplain water storage estimates with other sources such as those derived from routing algorithms. Satellite data should be used.

4.2 Impact of Hydraulic Parameters (ML and PTFDPM0)

Saturated hydraulic conductivity and the soil water retention curve parameters affect infiltration rate at the soil surface, soil moisture movement, recharge into groundwater, and streamflow generation, especially in timing and peak of the streamflow (ML's KGE). However, we could not observe significant effect of hydraulic parameter on generated BFI. Accurate estimation of these parameters can greatly improve streamflow predictions but is very challenging due to the complex soil texture, structure, and presence of coarse materials (See (Gupta, Papritz et al. 2022)) . ML-derived parameters showed a 20% improvement KGE, better matching observed streamflow patterns than traditional lookup tables used by the NWM. However, the limited geographic and climatic distribution of the training dataset contributing in generating ML parameters, may affect its generalization, potentially leading to biases in the predicted streamflow.

The PTFDPM0 scenario, which combines the dual permeability and the removal of ponding depth, showed a tendency to increase baseflow due to dual permeability while decreasing it because of the absence of a ponding threshold. This combination results in a reasonably accurate estimation of low BFI but underestimates high BFI. However, the inclusion of macropores in this setup helps to capture the highest BFI values, which other scenarios do not achieve (Figure 6b). Moreover, this configuration can reach the highest KGE at certain stations (Figure 9b). These findings highlight the importance of including both dual permeability and a ponding depth parameter to model baseflow accurately, particularly in regions with diverse hydrological conditions (Figure 3, 5, and 9).

4.3 Impact of Precipitation Datasets

The choice of precipitation datasets also significantly impacts the accuracy of baseflow predictions. Our results indicated that using the IMERG dataset improves the accuracy of BFI predictions in regions where NLDAS-2 tends to overestimate BFI. The IMERG experiments successfully captured lower BFI values in regions where the VGM configuration with AORC and NLDAS-2 precipitation overestimate them, particularly in the coastal regions of the California River Basin.

To better understand the effect of precipitation on baseflow generation, we analyzed the accumulated extreme values of precipitation as depicted in Figure S9, S10. Our analysis revealed that the effect of precipitation on baseflow can be categorized into two distinct groups. Firstly, increased precipitation intensity, characterized by heavy rainfall, enhances water infiltration into the soil. When the rainfall intensity is sufficiently high, it penetrates deeper soil layers and percolates into the groundwater. Consequently, as groundwater recharge intensifies, more baseflow is generated, leading to increased baseflow in regions marked by blue circles and green dashed lines in Figure S9, and S10. Conversely, when rainfall becomes excessively intense, it exceeds the soil's infiltration capacity. In such instances, water generates runoff rather than infiltrating into the soil and groundwater, thereby reducing baseflow in areas delineated by pink dashed lines (Figure S9, and S10). Overall, the impact of precipitation on baseflow varies by location and precipitation intensity. As precipitation increases, groundwater recharge and baseflow also increase; however, upon surpassing a threshold—likely related to soil infiltration capacity—groundwater discharge and baseflow should decrease. Thus, there is a need for detailed research on the impact of precipitation intensity on recharge and baseflow.

5 Conclusion

This study emphasizes the critical role of baseflow generation processes in streamflow prediction accuracy, especially in arid regions of the southwestern US. Our modeling results suggest that the streamflow prediction skill are sensitive to how baseflow generation is represented in terms of model physics, associated hydraulic parameters, and precipitation forcing data. Using a Noah-MP with enhanced hydrology, we show that the choice of hydrological processes, hydraulic parameters, and precipitation datasets significantly affects streamflow prediction accuracy over dry southwestern US.

The Van-Genuchten hydraulic scheme is more effective than the Brooks-Corey in modeling baseflow and BFI, particularly in dry regions where the soil is naturally dry, and the BFI is low. This scheme reduced the BFI overestimation produced by the Brooks-Corey with the CH hydraulic parameters (by the Noah-MP look-up table) and NWM with calibrated hydraulic parameters by better capturing groundwater recharge and discharge processes. Additionally, with the machine learning-derived soil water retention curve parameters, VGM significantly improves the streamflow predictions, offering a better match with the observed streamflow compared to the look-up table and pedotransfer functions. In general, our finding implies improving the baseflow in large-scale models like Noah-MP leads to better prediction of streamflow as observed in VGM configuration.

The study also highlights the importance of incorporating soil macropores, DPM experiment, and ponding depth thresholds, VGM and VGM0 experiments, in modeling, as these factors greatly influence infiltration, percolation, recharge and baseflow generation. A ponding depth greater than zero increases BFI by allowing more water to infiltrate, especially in wet regions. Additionally, the presence of macropores enhances drainage from the surface to the root zone, increasing baseflow and BFI. However, the benefits of these features vary by region. While uncalibrated macropore fraction improve predictions in high-BFI areas, they may lead to overestimations of baseflow in low-BFI regions.

Furthermore, the choice of precipitation dataset was shown to be crucial, with the IMERG dataset offering more accurate baseflow predictions in regions where traditional datasets like NLDAS-2 tended to overestimate BFI. Indeed, heavy precipitation facilitates the infiltration into deeper soil and groundwater recharge. This finding suggests that high-resolution precipitation data is essential for improving the accuracy of streamflow predictions in areas with complex hydrological conditions.

Overall, the study demonstrates that careful selection of hydrological processes (soil hydraulic schemes), hydraulic parameters, and precipitation datasets can significantly enhance the performance of hydrologic models in predicting streamflow, particularly in arid regions. These findings provide valuable insights for future research and model development, emphasizing the need to optimize model configurations before calibration to achieve more reliable streamflow predictions.

Appendix A

The Wösten et al. (1999) proposed Pedo-Transfer function:

$$\theta_s = 0.7919 + 0.001691 * C - 0.29619 * B - \ 0.000001491 * S^2 + 0.000082\,(SM)^2 + \frac{0.02427}{C} + \frac{0.01113}{S} + 0.01472 \\ * \ln(S) - 0.0000733 * (SM) * C - 0.000619 * B * C - 0.001183 * B * (SM) - 0.0001664 * top * S$$

$$\alpha = \exp\Big(-14.96 + 0.03135 * C + 0.0351 * S + 0.646 * (SM) + 15.29 * B - 0.192 * top - 4.671 * B^2 - 0.000781 * C^2 - 0.00687 * (SM)^2 + \frac{0.0449}{SM} + 0.0663 * \ln(S) + 0.1482 * \ln(SM) - 0.04546 * B * S - 0.4852 * B * SM + 0.00673 * topp * C\Big)$$

$$K_s = \exp\Big(7.75 + 0.352 * S + 0.93 * top - 0.967 * B^2 - 0.000484 * C^2 - 0.000322 * S^2 + \frac{0.001}{S} - \frac{0.0748}{SM} - 0.643 * \ln(S) - 0.01398 * B * C - 0.1673 * B * SM + 0.02986 * top * C - 0.03305 * top * S\Big)$$

$$n = 1 + \exp\Big(-25.23 - 0.02195 + 0.0074 * S - 0.1940 * SM + 45.5 * B - 7.24 * B^2 - 0.0003658 * C^2 + 0.002885 * (SM)^2 - \frac{12.81}{B} - \frac{0.1524}{S} - \frac{0.01958}{SM} - 0.2876 \ln(S) - 0.0709 \ln(SM) - 44.6 \ln(B) - 0.02264 * B * C + 0.0896 * B * SM + 0.00718 * top * C\Big)$$

$$l = 0.0202 + 0.0006193 * C^2 - 0.001136 * (SM)^2 - 0.2316 * \ln(SM) - 0.03544 * B * C + 0.00283 * B * S + 0.0488 * B * SM$$

$$L = \frac{10 * \exp(l) - 10}{\exp(l) + 1}$$

$$\theta_r = 0$$

Where C is Clay, S is Silt, SM is soil organic matter, BD is soil bulk density, and top is 1 for depths < 30 cm, otherwise 0.

Acknowledgments

Funding for this project was provided by the National Oceanic and Atmospheric Administration (NOAA), awarded to the Cooperative Institute for Research on Hydrology (CIROH) through the NOAA Cooperative Agreement with The University of Alabama, NA22NWS4320003. Also, the research carried out for this article was supported by the U.S. Army Corps of Engineers, Engineer Research and Development Center, Coastal Inlets Research Program via Congressionally Directed R&D with the National Oceanic and Atmospheric Administration's National Water Center.

The data used in this study are freely available online:
NLDAS-2 data (http://www.emc.ncep.noaa.gov/mmb/nldas/); NASA SMAP soil moisture product (https://nsidc.org/data/spl3smp_e/versions/6); GPM IMERG-Final product (https://disc.gsfc.nasa.gov/datasets/GPM_3IMERGHH_06/summary). The Noah-MP code used in this study has been uploaded to a repository that may be accessed by other researchers (https://github.com/mfarmani95/NoahMP_Dual).

Reference

Agnihotri, J., A. Behrangi, A. Tavakoly, M. Geheran, M. A. Farmani and G. Y. Niu (2023). "Higher Frozen Soil Permeability Represented in a Hydrological Model Improves Spring Streamflow Prediction From River Basin to Continental Scales." Water Resources Research **59**(4).

Arnold, J. G. and P. M. Allen (1999). "Automated Methods for Estimating Baseflow and Ground Water Recharge from Streamflow Records1." JAWRA Journal of the American Water Resources Association **35**(2): 411-424.

Bass, B., S. Rahimi, N. Goldenson, A. Hall, J. Norris and Z. J. Lebow (2023). "Achieving Realistic Runoff in the Western United States with a Land Surface Model Forced by Dynamically Downscaled Meteorology." Journal of Hydrometeorology **24**(2): 269-283.

Batalha, M. S., M. C. Barbosa, B. Faybishenko and M. T. Van Genuchten (2018). "Effect of temporal averaging of meteorological data on predictions of groundwater recharge." Journal of Hydrology and Hydromechanics **66**(2): 143-152.

Beck, H. E., A. I. J. M. van Dijk, D. G. Miralles, R. A. M. de Jeu, L. A. Sampurno Bruijnzeel, T. R. McVicar and J. Schellekens (2013). "Global patterns in base flow index and recession based on streamflow observations from 3394 catchments." Water Resources Research **49**(12): 7843-7863.

Bisht, G., W. J. Riley, G. E. Hammond and D. M. Lorenzetti (2018). "Development and evaluation of a variably saturated flow model in the global E3SM Land Model (ELM) version 1.0." Geosci. Model Dev. **11**(10): 4085-4102.

Blöschl, G., M. Sivapalan, T. Wagener, A. Viglione and H. Savenije (2013). Runoff Prediction in Ungauged Basins.

Burnash, R. (1995). "The NWS River Forecast System-catchment modeling."

Chapman, T. (1999). "A comparison of algorithms for stream flow recession and baseflow separation." Hydrological Processes **13**(5): 701-714.

Chapman, T. G. (1991). "Comment on “Evaluation of automated techniques for base flow and recession analyses” by R. J. Nathan and T. A. McMahon." Water Resources Research **27**(7): 1783-1784.

Chen, S. and X. Ruan (2023). "A hybrid Budyko-type regression framework for estimating baseflow from climate and catchment attributes." Journal of Hydrology **618**.

Cosgrove, B., D. Gochis, T. Flowers, A. Dugger, F. Ogden, T. Graziano, E. Clark, R. Cabell, N. Casiday, Z. Cui, K. Eicher, G. Fall, X. Feng, K. Fitzgerald, N. Frazier, C. George, R. Gibbs, L. Hernandez, D. Johnson, R. Jones, L. Karsten, H. Kefelegn, D. Kitzmiller, H. Lee, Y. Liu, H. Mashriqui, D. Mattern, A. McCluskey, J. L. McCreight, R. McDaniel, A. Midekisa, A. Newman, L. Pan, C. Pham, A. RafieeiNasab, R. Rasmussen, L. Read, M. Rezaeianzadeh, F. Salas, D. Sang, K. Sampson, T. Schneider, Q. Shi, G. Sood, A. Wood, W. Wu, D. Yates, W. Yu and Y. Zhang (2024). "NOAA's National Water Model: Advancing operational hydrology through continental-scale modeling." JAWRA Journal of the American Water Resources Association **60**(2).

Crosbie, R. S., J. L. McCallum, G. R. Walker and F. H. Chiew (2012). "Episodic recharge and climate change in the Murray-Darling Basin, Australia." Hydrogeology Journal **2**(20): 245-261.

David, C. H., D. R. Maidment, G.-Y. Niu, Z.-L. Yang, F. Habets and V. Eijkhout (2011). "River Network Routing on the NHDPlus Dataset." Journal of Hydrometeorology **12**(5): 913-934.

de Roo, A., H. E. Beck and A. I. J. M. van Dijk (2015). "Global Maps of Streamflow Characteristics Based on Observations from Several Thousand Catchments*." Journal of Hydrometeorology **16**(4): 1478-1501.

Dourte, D., S. Shukla, P. Singh and D. Haman (2013). "Rainfall intensity-duration-frequency relationships for Andhra Pradesh, India: changing rainfall patterns and implications for runoff and groundwater recharge." Journal of hydrologic Engineering **18**(3): 324-330.

Eckhardt, K. (2008). "A comparison of baseflow indices, which were calculated with seven different baseflow separation methods." Journal of Hydrology **352**(1-2): 168-173.

Fall, G., D. Kitzmiller, S. Pavlovic, Z. Zhang, N. Patrick, M. St. Laurent, C. Trypaluk, W. Wu and D. Miller (2023). "The Office of Water Prediction's Analysis of Record for Calibration, version 1.1: Dataset description and precipitation evaluation." JAWRA Journal of the American Water Resources Association **59**(6): 1246-1272.

Farmani, M. A., A. Behrangi, A. Gupta, A. Tavakoly, M. Geheran and G. Y. Niu (2024). "What Are the Key Soil Hydrological Processes to Control Soil Moisture Memory?" EGUsphere **2024**: 1-28.

Furey, P. R. and V. K. Gupta (2001). "A physically based filter for separating base flow from streamflow time series." Water Resources Research **37**(11): 2709-2722.

Genereux, D. (1998). "Quantifying uncertainty in tracer-based hydrograph separations." Water Resources Research **34**(4): 915-919.

Ghimire, G. R., C. Hansen, S. Gangrade, S. C. Kao, P. E. Thornton and D. Singh (2023). "Insights From Dayflow: A Historical Streamflow Reanalysis Dataset for the Conterminous United States." Water Resources Research **59**(2).

Gupta, H. V., H. Kling, K. K. Yilmaz and G. F. Martinez (2009). "Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling." Journal of Hydrology **377**(1-2): 80-91.

Gupta, S., A. Papritz, P. Lehmann, T. Hengl, S. Bonetti and D. Or (2022). "Global Mapping of Soil Water Characteristics Parameters— Fusing Curated Data with Machine Learning and Environmental Covariates." Remote Sensing **14**(8).

Hall, F. R. (1968). "Base-Flow Recessions—A Review." Water Resources Research **4**(5): 973-983.

Hansen, C., J. S. Shiva, S. McDonald and A. Nabors (2019). "Assessing Retrospective National Water Model Streamflow with Respect to Droughts and Low Flows in the Colorado River Basin." Journal of the American Water Resources Association **55**(4): 964-975.

Hengl, T., J. M. de Jesus, R. A. MacMillan, N. H. Batjes, G. B. Heuvelink, E. Ribeiro, A. Samuel-Rosa, B. Kempen, J. G. Leenaars, M. G. Walsh and M. R. Gonzalez (2014). "SoilGrids1km--global soil information based on automated mapping." PLoS One **9**(8): e105992.

Hong, Y., H. Xuan Do, J. Kessler, L. Fry, L. Read, A. Rafieei Nasab, A. D. Gronewold, L. Mason and E. J. Anderson (2022). "Evaluation of gridded precipitation datasets over international basins and large lakes." Journal of Hydrology **607**.

Huang, J., P. Wu and X. Zhao (2013). "Effects of rainfall intensity, underlying surface and slope gradient on soil infiltration under simulated rainfall experiments." Catena **104**: 93-102.

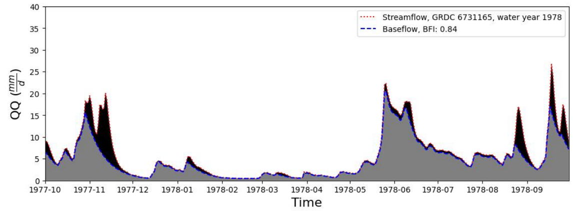

Figure S1. Baseflow computed using the method suggested by Van Dijk (2010) for GRDC 6731165 station data to compare with Beck et al (2013).

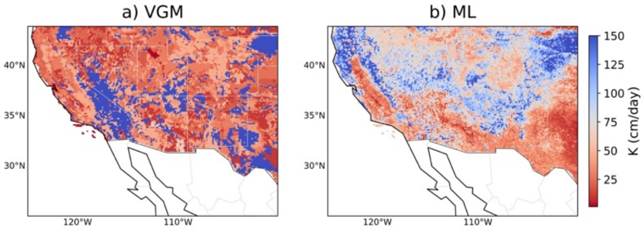

Figure S2. Hydrohalic conductivity of a) VGM and b) ML scenarios. ML shows higher hydraulic conductivit over most of regions.

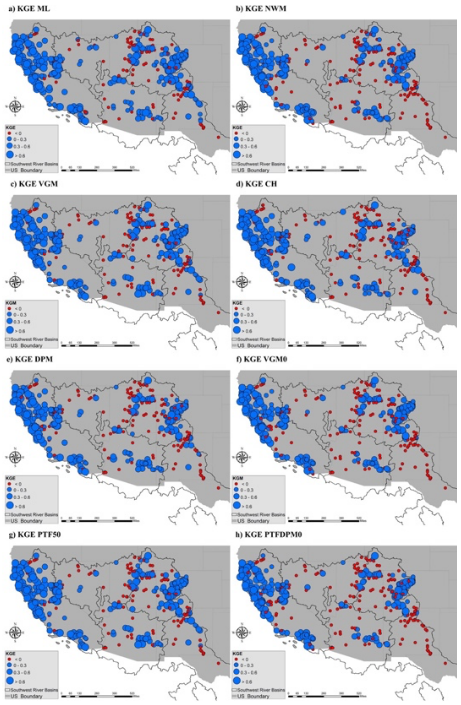

Figure S3. KGE computed for 1980–2019 for (a) ML, (b) NWM, (c) VGM, (d) CH, (e) DPM, (f) VGM0, (g) PTF50, and (PTFDPM0. Negative KGE values are seen in high BFI regions (Great Basin, Upper Colorado, Rio Grande), while positive values are found in the California River Basin. Some Rio Grande and Upper Colorado stations show negative KGE for NWM but positive for ML.

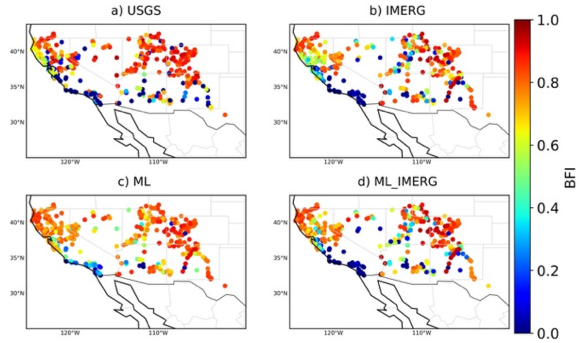

Figure S4. BFI at 390 locations across the southwestern region. Panels show the following: (a) USGS; (b) IMERG; (c) ML; and (d) ML_IMERG which is combination of ML and IMERG scenarios. The BFI for all scenarios is computed for 2014-2019 to be consistent with forcing scenarios. Combining IMERG and ML could help to better capture both low and high BFI than IMERG and ML scenarios. Hence, ML_IMERG outperform eths ML and IMERG in simulating the baseflow.

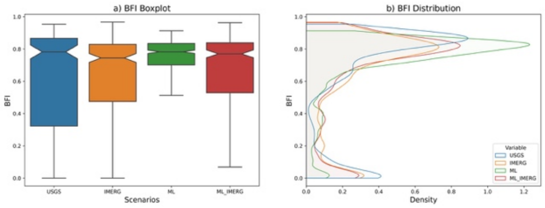

Figure S5. BFI at 390 locations across the southwestern region. Panels display: (a) BFI boxplot for different scenarios and (b) BFI distribution for USGS, IMERG, ML, and ML_IMERG, which combines ML and IMERG scenarios. The BFI for all scenarios is computed for 2014–2019 to maintain consistency with the forcing data. Combining IMERG and ML helps capture both low and high BFI (median) more effectively than either IMERG or ML alone, with ML_IMERG outperforming both in simulating baseflow.

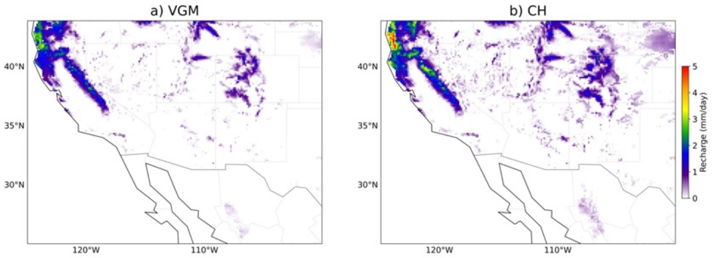

Figure S6. Average Noah-MP recharge over 1980-2019 for (a) VGM and (b) CH scenarios.

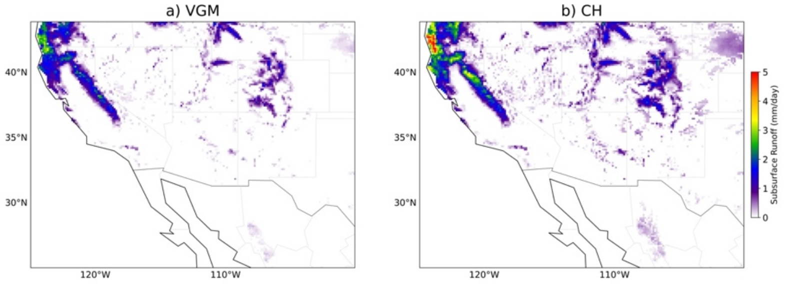

Figure S7. Average Noah-MP subsurface runoff over 1980-2019 for (a) VGM and (b) CH scenarios.

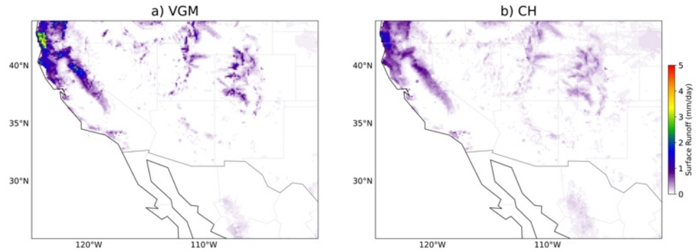

Figure S8. Average Noah-MP surface runoff over 1980-2019 for (a) VGM and (b) CH scenarios.

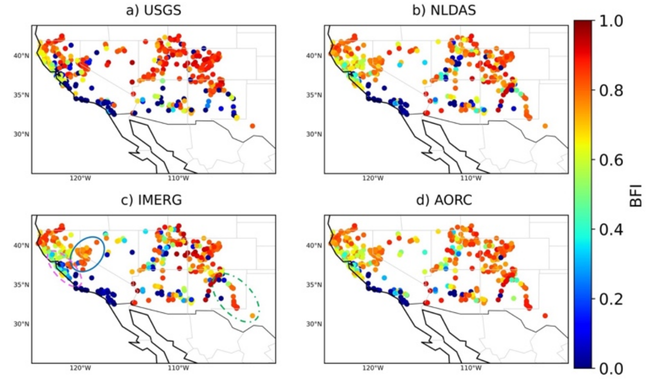

Figure S9. This figure is the same as the figure 7 in paper. We put the plot here to specify the regions affected by the precipitation intensity. The specifies locations are to demonstrate how precipitation affect the BFI. Check Figure S10 to better understand relation between heavy precipitation and BFI.

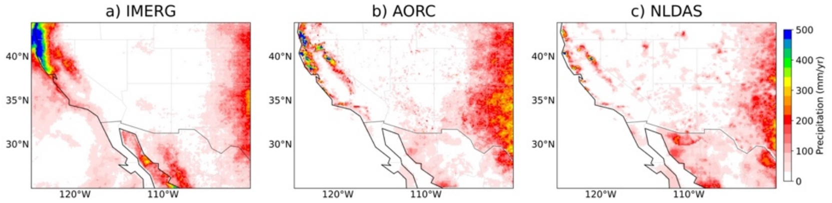

Figure S10. Accumulated annual extreme precipitation (>10mm/hr) (a) IMERG, (b) AORC, and (c) NLDAS-2.