A Calibration-Free Groundwater Module for Improving Predictions of Low Flows

Arik Tashie[1], Tamlin Pavelsky[2], and Mukesh Kumar[1]

[1] Department of Civil, Construction, and Environmental Engineering, University of Alabama, AL, USA.

[2] Department of Geological Sciences, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA.

Corresponding author: Arik Tashie (tashi002@ua.edu)

Key Points:

- Develop a calibration-free groundwater / low flow module that is compatible with Land Surface Models (LSMs) and rainfall-runoff models.
- Without calibration, the module's performance compares favorably with an existing, fully calibrated alternative.
- Module predictions of low flows are robust to drastic changes in the overlying hydrologic model.

Abstract

Groundwater modules are critically important to the simulation of low flows in land surface models (LSMs) and rainfall-runoff models. Here, we develop a Groundwater for Ungauged Basins (GrUB) module that uses only physically-based properties for which data are widely available, thus allowing its application without the need for calibration. GrUB is designed to be computationally simple and readily adaptable to a wide variety of LSMs and rainfall-runoff models. We assess the performance of GrUB in 84 US watersheds by incorporating it into HBV, a popular rainfall-runoff model. We compare predictions of low flows by the native (calibrated) HBV groundwater module with those by the (uncalibrated) GrUB module and find that GrUB generates error metrics that are equivalent to (or superior to) those generated by the native HBV groundwater module. To assess whether predictions by GrUB are robust to changes in the structure and parameterization of the overlying hydrologic model, we run tests for two artificial scenarios: Slow Recharge with rates of percolation below 0.1 mm/day, and Fast Recharge with rates of percolation of up to 1000 mm/day. GrUB proves to be robust to these extreme changes, with mean absolute error (MAE) of predictions of low flows only increasing by an average of up to 19%, while average MAE increases by up to 157% when the same tests are performed on HBV without the GrUB module. We suggest GrUB as a potential tool for improving predictions of low flows in LSMs as well as rainfall-runoff models where calibration data are unavailable.

1. **Introduction**
   (a) **Background**

Shallow groundwater supplies a majority of streamflow to most watersheds [Beck et al. 2013] and is the primary source of streamflow during seasonal low flows [Smakhtin 2001]. This mobile water resource provides an essential buffer to changes in temperature, nutrients, and precipitation [Ficklin et al. 2015] and sustains evapotranspiration (ET) [Yang et al. 2011] during low flow periods. Even in snow-dominated climates, the hydrologic response to snowmelt is mediated by groundwater [Enzminger et al. 2019] with important implications for the response of these systems to global warming [Tague and Grant [2009]. Because shallow groundwater contributions to low flows are disproportionately sensitive to changes in near-term (i.e., years and decades) climate signals [Hare et al. 2021], it is essential that our hydrologic models are able to predict groundwater contributions to low flows accurately. Unfortunately, accurate simulation of low flows and groundwater contributions to them has proven to be notoriously difficult in land surface models (LSMs) [Clark et al. 2015, Holtzman 2020] and conceptual rainfall-runoff models [Fowler et al. 2019].

Early LSMs described low flows as dependent on one-dimensional drainage below a soil column [Clark et al. 2015]. This "free" drainage led to well-documented inaccuracies: too-fast drainage during wet periods, underestimates of seasonal storage, and the cessation of ET during short dry periods [Baker et al. 2008, Brunke et al. 2016, Fan et al. 2017, Kuppel et al. 2017, Milly and Shmakin 2002, Miguez-Macho and Fan 2012a, Miguez-Macho and Fan 2012b, Pokhrel et al. 2013]. Though LSM researchers have introduced several mechanisms for constraining low flows, this additional model complexity [NOAA 2016] has failed to generate consistent improvements in model predictions [Yang et al. 2011, Gan et al. 2019]. Indeed, it is often necessary to incorporate a calibrated groundwater module in LSMs [Fang et al. 2019, Holtzman et al. 2020] when using LSM projections in a predictive context. The requirement of calibration both increases the complexity of applying LSMs and (importantly) limits their utility in ungauged basins.

The failure to clearly improve low flows in LSMs has been partially attributed to three issues. First, the simple (linear) groundwater reservoirs [e.g., Liang et al. 2003 or Niu et al. 2007] and quasi-TOPMODEL modules [e.g. Niu et al. 2005 or Oleson et al. 2010] may lack the complexity to characterize heterogeneous watershed features that drive highly nonlinear streamflow processes during periods of recession [Clark et al. 2009, Clark et al. 2015, Fan et al. 2019, Rahman et al. 2019, Tashie et al. 2019]. Second, the data used to parameterize these low flow modules are highly uncertain. For example, the hydraulic properties of the deep subsurface are typically estimated according to the texture of the shallow overlying soil using a pedotransfer function [Gedney and Cox 2003]. Even where it is appropriate to assume a strong correlation between hydraulic properties of shallow soils and those of the deeper soils, regolith, and bedrock, this approach ignores macropores [Zecharias and Brutsaert 1988, Mendoza et al. 2004] and is biased in non-temperate climates [Hengl et al. 2017, Huscroft et al. 2018]. Finally, the overreliance on only one or two datasets (i.e., areal average hydraulic conductivity and/or topographic wetness index) to character-

ize a complex process heightens the likelihood that the biases and uncertainties that are inherent in all large datasets express themselves in model predictions [Hariri et al. 2019].

These issues are not limited to LSMs, but are also increasingly being recognized as an area of needed improvement in conceptual rainfall-runoff models [Fowler et al. 2019, Seibert and van Meerveld 2016]. Computationally simple rainfall-runoff models are commonly applied in climate change impact studies to inform policy and decision-making in a wide variety of arenas [Flörke et al., 2018; Iqbal et al., 2018, Mahmoud & Gan, 2018, Cui et al., 2018, Balkovič et al., 2018, de Jong et al., 2018, Emanuel, 2018]. However, rainfall-runoff models tend to become increasingly inaccurate during periods of hydroclimatic variability [Saft et al. 2016, Seibert and van Meerveld 2016], which inhibits their utility in predicting watershed response to climate. Further, most rainfall-runoff models require calibration of several parameters on a single objective function, which raises the issue of equifinality [Beven 2006] and limits their utility in ungauged and poorly gauged basins [Boughton 2006].

1. **Objectives**

We propose to improve predictions of low flows in LSMs and conceptual rainfall-runoff models by developing a portable, data-driven module called Groundwater for Ungauged Basins (GrUB). To facilitate easy and broad adoption of this module by the modeling community, we are guided by four key practical principles:

1. No calibration required: the module must be usable directly "out of the box" and easily applied in ungauged basins.
2. Simple data requirements: the module must depend only on widely available (continental- or global-scale) data that requires minimal processing on the part of the model user.
3. Modular: to be adaptable to a variety of LSMs and rainfall-runoff models, the module must be driven by a single flux that is common in most LSMs and rainfall-runoff models (i.e., deep recharge), and it must operate independently of the rest of the model structure (i.e., no feedback mechanisms).
4. Computational simplicity: the module must not substantially increase the run time of LSMs (which are already computationally complex) or rainfall-runoff models (for which short run times are a major source of appeal).

Following the module development (detailed below), we incorporate GrUB into a common rainfall-runoff model (HBV) and assess its performance by posing the following questions:

1. Does incorporating the uncalibrated GrUB module into the otherwise calibrated HBV model negatively affect its overall performance?
2. Does GrUB reduce error and bias in predictions of low flows, especially during historically long dry periods?

3. Are GrUB predictions of low flows robust to changes in the parameterization and performance of the overlying hydrologic model?

Finally, we discuss the limitations of GrUB and identify potential avenues for improvement.

1. **Model Development**

   (a) **Conceptual Model**

In individual watersheds (or hillslopes), actual groundwater flow patterns are exceedingly complex, varying in time, place, and with antecedent conditions [Aulenbach et al. 2021, Tashie et al. 2019, Zimmer and McGlynn 2017a, Zimmer and McGlynn 2017b]. Which mechanisms are dominant (and when) varies according to the specific geophysical properties and climatic fluxes of a watershed [Tashie et al. 2020b]. Unfortunately, the practical realities of a large-scale hydrologic and land surface modeling demand that low flow calculations must be generic (i.e., the same for all watersheds) and computationally efficient.

In this research, we apply a bottom-up, data driven approach to module development. Broadly, we propose that models that leverage a wide variety of empirical data are more likely to be transferrable among watersheds without the need for calibration. We begin by identifying a class of conceptual models that incorporates a wide variety of physically-based properties and mechanisms for which extensive empirical data is readily available. Then, we build a module that is simple and generic, so that it is applicable across a range of watersheds using only these widely available data.

Representation of groundwater contributions to low flows in LSMs is usually based on one of four classes of conceptual models [Clark et al. 2015]: 1) free or restricted drainage; 2) conceptual buckets; 3) TOPMODEL; and 4) hillslope flow based on Darcy's Law. Free drainage and conceptual buckets generate subsurface runoff (Qsb) as a function of storage (S) and saturated hydraulic conductivity (Ksat):

$$Q_{\mathrm{sb}} = f(S,\ K) \quad \text{Eq. 1}$$

where S is a function of modeled inputs to and outputs from the lower boundary and K is usually derived from the properties of lowest layer of the soil column. Free and restricted drainage modules were the most common approaches in early LSMs and remain popular (e.g., CABLE, TESSEL, Noah, and ORCHIDEE) [Kowalczyk et al. 2013 , van den Hurk et al. 2000, Niu et al. 2011, and Krinner et al. 2005].

To address the poor performance of free and restricted drainage models, many LSMs have incorporated TOPMODEL concepts (e.g., Catchment, CLM, JULES, MATSIRO, and Noah-MP), where Q_{sb} is again a function of S and K, but also depends on catchment-average topographic slope and curvature, i.e., the topographic wetness index (TWI):

$$Q_{\mathrm{sb}} = f(S,\ K, TWI) \quad \text{Eq. 2}$$

In TOPMODEL approaches, S is again a function of modeled inputs to and outputs from the lower boundary, while TWI and K are largely derived from empirical data. TWI is calculated directly from digital elevation models (DEMs) and K varies with S, such that maximum K is derived from the properties of the lowest layer of the soil column and decays exponentially with declining values of S [e.g., Niu et al. 2005].

Finally, flow from a representative hillslope based on Darcy's Law has been explicitly identified as a potential mechanism for improving representation of groundwater and low flows [Clark et al. 2015, Fan et al. 2019], though only one version of this mechanism has been included in any LSM to date (i.e., LM3 [Subin et al. 2014]). Broadly, a representative hillslope flow model describes Qsb as a function of S, K, and several topographic and geomorphic variables:

$Q_{\mathrm{sb}} = f(S,\ K, i,\ B,\ L, Th)$ Eq. 3

where i is aquifer slope, B is characteristic aquifer breadth, L is stream network length, and Th is aquifer thickness. Once again, S is a function of modeled inputs to and outputs from the lower boundary. K, i, B, L, and Th represent physical properties of the watershed for which empirical data is widely available. These empirical data represent an opportunity to uniquely parameterize watersheds without calibration, and without overreliance on any single dataset or set of simplifying assumptions that may be prone to bias. Therefore, we proceed to develop the GrUB module according to this approach.

1. **Model Equations**

We begin with a continuity equation and add complexity where hydrological realism demands further refinement:

$S = N - \ Q_{\mathrm{sb}}$ Eq. 4

where N is recharge, i.e. modeled drainage from the unsaturated zone (Figure 1). Q_{sb} [L^3] is a product of linear discharge from the hillslope (q_{sb}), the thickness of the saturated aquifer at the stream interface (H_0), and the total length of stream network (L):

$Q_{\mathrm{sb}} = q_{\mathrm{sb}} H_0 L = K \frac{\mathrm{dh}}{\mathrm{dx}} H_0 L$ Eq. 5

where $\frac{\mathrm{dh}}{\mathrm{dx}}$ is the hydraulic gradient. Watershed-scale estimates of L can be derived from topography, remote sensing, and modeling experiments, and these are data widely available in public databases like the National Hydrography Dataset (NHD) [USGS 2004]. Because the 1:100,000-scale stream networks in the NHD tend to underestimate actual L by a factor of at least 2 [Godsey and Kirchner 2014, Jensen et al 2015], we rescale L by a factor of 2 for all watersheds. This leaves K, $\frac{\mathrm{dh}}{\mathrm{dx}}$, and H_0 to be solved for.

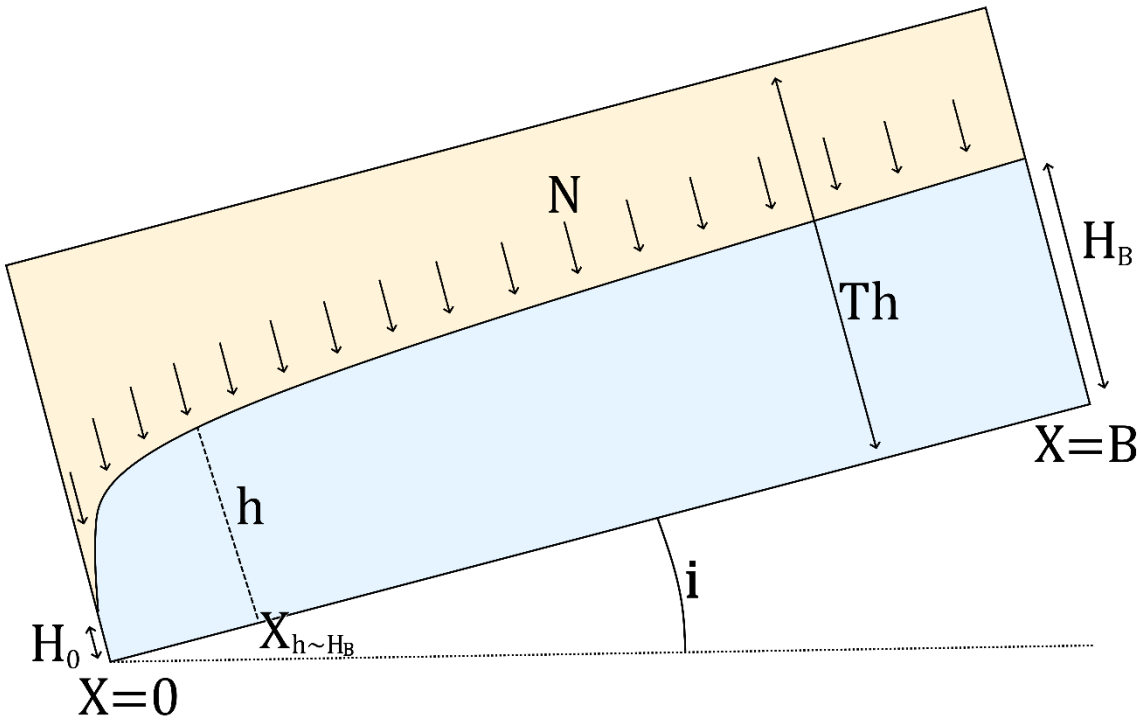

Figure 1: A representative hillslope aquifer atop a sloping impermeable layer.

As noted above, K in LSMs is typically estimated according to the soil texture at the bottom of the soil column according to a pedotransfer function and modeled to decay exponentially with depth to water table [Niu et al. 2005]. However, pedotransfer functions are known to be highly uncertain [Freeze and Cherry 1979, Zhang and Schaap 2019] and to systematically underestimate K by ignoring macropores [Zecharias and Brutsaert 1988, Mendoza et al. 2004]. Further, the exponential decay of K is a purely empirical relationship that is often calibrated in hydrologic models [e.g., Tague and Band 2004]. To address these pitfalls, Tashie et al. [2021] developed maps of a quantity we term effective K for the coterminous US that: 1) accounts for the effects of macropores on watershed-average values of effective K; and 2) explicitly represents values of K as a function of shallow groundwater storage (S). Using these data, we can solve K as a function of (static) watershed properties and (time-varying) S:

$$\log_{10}(K) = \log_{10}(K_0) + Sm \quad \text{Eq.7}$$

Here, K_0 is the value of K during extremely dry conditions ($S0$) and m is an empirically derived constant unique to each watershed (calculated according to Tashie et al. [2021]). This dataset also has the benefit of providing an estimate

of maximum saturated aquifer thickness (Th) which is useful for approximating the two remaining terms to be solved: $\frac{\mathrm{dh}}{\mathrm{dx}}$ and H_0.

Extensive literature exists deriving analytical solutions to $\frac{\mathrm{dh}}{\mathrm{dx}}$ and H_0 as functions of Th, S, N, and drawdown time, with each set of solutions applying its own set of assumptions and approximations that nonetheless broadly agree in describing the dynamic response of the hillslope aquifer to recharge and drainage [see Troch et al. 2013]. The water table elevation profile ($h(x)$) of a fully recharged hillslope aquifer is "flat" ($H_0 \cong Th \cong H_B$). During drainage, it drains more quickly nearer the stream ($X \to 0$) generating a curvilinear $h(x)$ profile that decays exponentially from $X = B$ to $X = 0$. The effective hydraulic gradient of a hillslope aquifer on a non-sloped ($i = 0$) impermeable layer may be approximated as the maximum hydraulic head within the hillslope (H_B) at the minimum distance from the stream where $h(X) \sim H_B$ (*i.e.*, $X_{h\sim B}$). This is illustrated in Figure 2, where $X_{h\sim B}$ is delineated as the location (X) before which there is a sharp decrease in hydraulic head (h). The total hydraulic gradient of a hillslope aquifer on a sloped ($i > 0$) impermeable layer may be approximated as:

$$\frac{\mathrm{dh}}{\mathrm{dx}} \cong \frac{H_B + X_{h\sim H_B} \sin i}{X_{h\sim H_B}} \quad \text{Eq.8}$$

Though there is no universal method for calculating $X_{h\sim H_B}$ following extended periods of drawdown, it may be approximated as some fraction of the distance $X = B$ according to the ratio of maximum aquifer storage ($S_{\max} = \frac{\mathrm{Th}}{f}$) and the current aquifer storage deficit ($S_{\max} - S$) to instantaneous storage (S):

$$X_{h\sim H_B} \cong B\left(\frac{S_{\max} - S}{S_{\max}}\right)^w \quad \text{Eq.9}$$

where $S_{\max}$ and w are empirically derived constants, and f is drainable porosity. Calculations of $S_{\max}$ are available for all watersheds in the coterminous US [Tashie et al. 2021], though w must be estimated. We make a universal estimate of w=3 for all watersheds based on the following practical guidelines: 1) in actual hillslopes, an aquifer that is "nearly saturated" (>90%) is likely to approach a unit gradient; 2) median B for the coterminous US is >1000m, as estimated using stream density data from the NHD [USGS 2004]; 3) median Th for the coterminous US is 3m [Tashie et al. 2021]; and 4) for a "nearly saturated" watershed, $\frac{\mathrm{dh}}{\mathrm{dx}} \cong \frac{H_B}{X_{h\sim H_B}} \cong \frac{H_B}{B\left(\frac{S_{\max} - S}{S_{\max}}\right)^w} \cong \frac{(3*0.9)}{1000(.1)^w}$. w=3 is the whole number which most closely approximates $\frac{\mathrm{dh}}{\mathrm{dx}}$=1. We provide sensitivity analysis for values of w in section 4.4.

We take a similarly practical approach to approximating H_0. H_0 increases linearly with N immediately following a pulse of recharge, such that initial conditions of a dry aquifer ($S = 0$) following a pulse of recharge (N) are $H_0 \cong N \cong H_B$, after which the effect of N on H_0 decays exponentially. Therefore, at any time t:

$$H_0 \cong \sum_{t=0}^{-\infty} \frac{N_t}{f} v^{-t} \quad \text{Eq.10}$$

where N_t is total recharge during the previous time step t (e.g., t days ago) and v is an empirically derived constant. Because LSMs and rainfall-runoff models are solved iteratively, Eq.10 can be simplified for computational efficiency as:

$H_0 = \frac{N_0}{f} + H_{0_{t=-1}} v$ Eq.11

where $H_{0_{t=-1}}$ is equal to H_0 during the previous time step. To account for well-documented seasonal hysteresis in low flow recession signatures [Shaw and Riha 2012, Bart and Hope 2014], we selected a value of v such that N from more than three months previous ($t <$ -90*)* has a negligible effect on H_0 ($v^{90} < 0.01$). Therefore, we set $v = 0.95$ for all watersheds, and provide sensitivity analysis for this variable in section 4.4. However, H_0 does not reach 0 during periods without recharge, but instead asymptotically approaches some minimum value relative to H_B. We estimated a minimum value of H_0 relative to H_B according to the following criteria: 1) H_0 linearly covaries with H_B, such that $H_{\text{minimum}} = \frac{H_B}{p}$ where p is an empirical constant; 2) in actual hillslopes, a reasonable minimum thickness of the stream-aquifer interface is likely to be on the scale of (tens of) centimeters; 3) the median *Th* for the coterminous US is 3m [Tashie et al. 2021]; and 4) therefore *p=100* provides a realistic minimum constraint on $H_{minimum}$ relative to H_B. Thus:

$H_0 = \frac{N_0}{f} + H_{0_{t=-1}} v + \frac{H_B}{p}$ Eq.12

Solving Eq. 5 with Eq. 7, Eq. 8, Eq. 9, and Eq. 12 gives:

$$Q_{\text{sb}} = (K_0 + 10^{\text{Sm}}) \left(\frac{\left(\frac{S}{f}\right)}{B\left(\frac{S_{\text{max}} - S}{S_{\text{max}}}\right)^w} + \sin i \right) \left(\frac{N_0}{f} + H_{0_{t=-1}} v + \frac{\left(\frac{S}{f}\right)}{p} \right) \text{L} \quad Eq.13$$

This computationally efficient model structure contains 7 parameters that are watershed-scale empirical values (K_0, m, S_{max}, B, f, L, and i), 3 parameters that are universal estimates (w, p and v), 1 variable that is supplied by the overlying LSM or rainfall-runoff model (N_0), and 2 variables that are solved iteratively by the GrUB low flow module (S and $H_{0_{t=-1}}$). As noted above, we provide sensitivity analysis for p, v, and w in section 4.4 because their values are neither analytically derived nor empirically estimated.

1. **Model Testing**

   (a) **Description of HBV**

An initial assessment of the potential uncertainties and biases of the GrUB module necessitates analysis of a variety of flow conditions (i.e., long time periods) across a variety of physioclimates (i.e., a large number of basins). We chose to incorporate GrUB within the HBV model [Bergström 1976, Bergström and Lindström 2015] because of its modular structure that is parsimonious yet adaptable to a wide range of physical and climatic settings [Bergstrom 1992]. Specifically, we adapted HBV.IANIGLA [Toum 2021] which interfaces with R software ["R Core Team" 2019]. Apart from its appeal as being computationally simple and broadly applicable, we chose to implement HBV.IANIGLA because the subroutines it uses to generate streamflow are conceptually similar to those

used in Noah-MP and other common LSMs (Figure 2). Here, we give a brief outline of HBV model structure, though for a detailed description we direct readers to Bergström [1976], Bergström [1992], and Bergström and Lindström [2015].

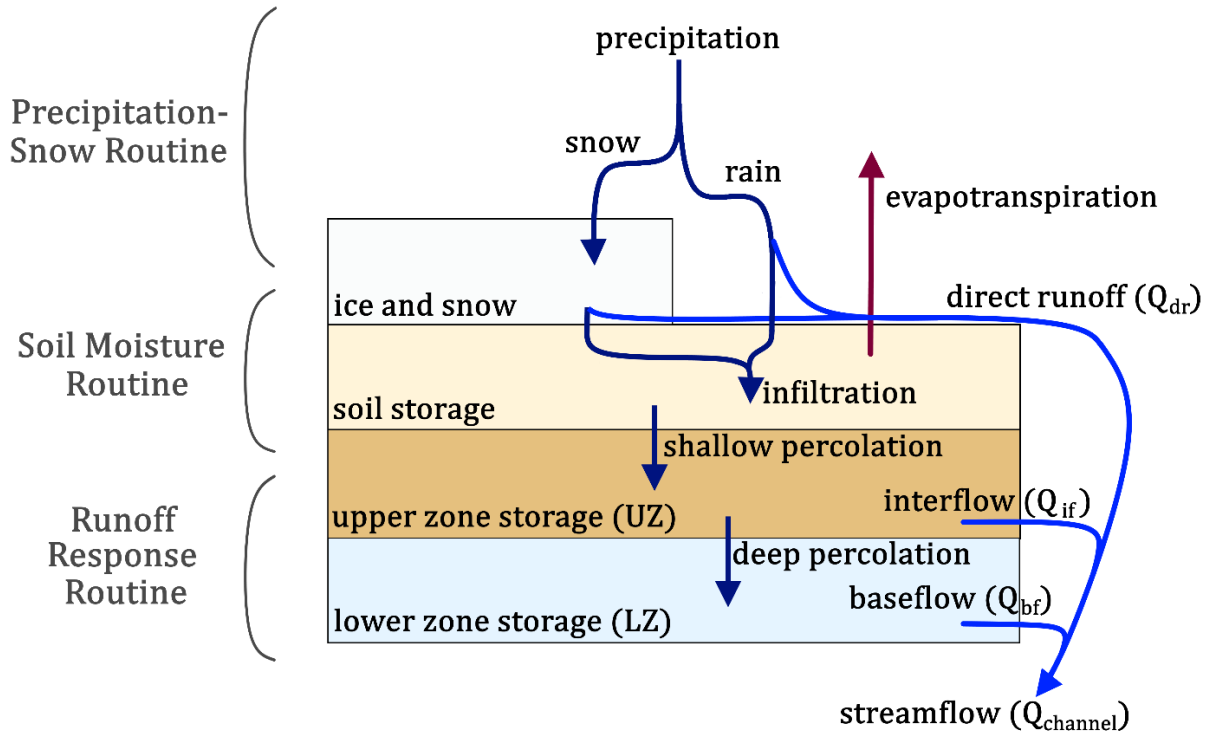

Figure 2: Conceptual model of HBV.

HBV operates as a sequence of subroutines with parameters within each subroutine requiring calibration [Sælthun 1996] (Figure 2). HBV.IANIGLA requires two time series of climate data (mean temperature (T_{mean}) and precipitation (PPT)) to drive the model, and a single time series (generally streamflow (Q)) for calibration. The first routine ("Precipitation-Snow Routine") calculates rainfall, snowfall, snowmelt, and potential evapotranspiration (PET) from T_{mean} and PPT. The second routine ("Soil Moisture Routine") calculates actual evapotranspiration (AET) and infiltration according to inputs from the first routine. The third and final routine ("Runoff Response Routine") calculates direct runoff (Q_{dr}), interflow (Q_{if}), and baseflow (Q_{bf}) according to inputs from the first two routines. The three runoff components are each independently calculated according to linear discharge from each of three conceptual buckets, with total streamflow ($Q_{channel}$) calculated as their sum. The parameters requiring calibration for each subroutine are listed in Table 1.

Table 1: parameters used to calibrate HBV

Parameter name	Parameter description	Routine	Minimu
SFCF	snowfall correction factor [-]	Precipitation-Snow Routine	.2
TR	solid-liquid PPT threshold temperature [C]	Precipitation-Snow Routine	-6
TT	melt temperature [C]	Precipitation-Snow Routine	-6
FM	snowmelt factor [mm/C]	Precipitation-Snow Routine	.2
FI	icemelt factor [mm/C]	Precipitation-Snow Routine	.2
FIC	debris-covered icemelt factor [mm/C]	Precipitation-Snow Routine	2
FC	soil field capacity [mm]	Soil Moisture Routine	25
LP	AET correction factor [-]	Soil Moisture Routine	.2
Beta	soil storage-runoff exponential [-]	Soil Moisture Routine	1
K0	top bucket (STZ) storage constant [1/t]	Runoff Response Routine	.05
K1	intermediate bucket (SUZ) storage constant [1/t]	Runoff Response Routine	.005
K2	lower bucket (SLZ) storage constant [1/t]	Runoff Response Routine	.0001
UZL	max flux rate between STZ and SUZ [mm/t]	Runoff Response Routine	.2
Perc	max flux rate between SUZ and SLZ [mm/t]	Runoff Response Routine	.1

A benefit of the minimal data requirements of HBV is that a large number of watersheds are available for model calibration and assessment. We chose to use data from the Model Parameter Estimation Experiment (MOPEX) data set [Schaake et al. 2006] due to their strict data standards. Of the 438 MOPEX basins for which hydrometeorological data are available, we selected 84 based on: 1) being designated as minimally impacted by human interference in the GAGES-II dataset [Falcone et al. 2010]; and 2) having a minimum of 20 years of unflagged streamflow data. These watersheds are illustrated in Figure 3.

For each watershed, we calibrated the 14 parameters from Table 1 using the R-software package hydroGOF [Zambrano-Bigiarini 2020]. We used a Monte-Carlo approach, generating 5000 random samples from a uniform distribution of parameter values as listed in Table 1. We removed the first 3 years of data from each watershed before calculating the Kling-Gupta efficiency (KGE) [Gupta et al. 2009] scores, to allow the groundwater stores to equilibrate. Model parameterization was chosen based optimal KGE calibration; because model *performance* was not assessed with KGE but instead with low flow metrics (see below) no data were reserved for validation using KGE.

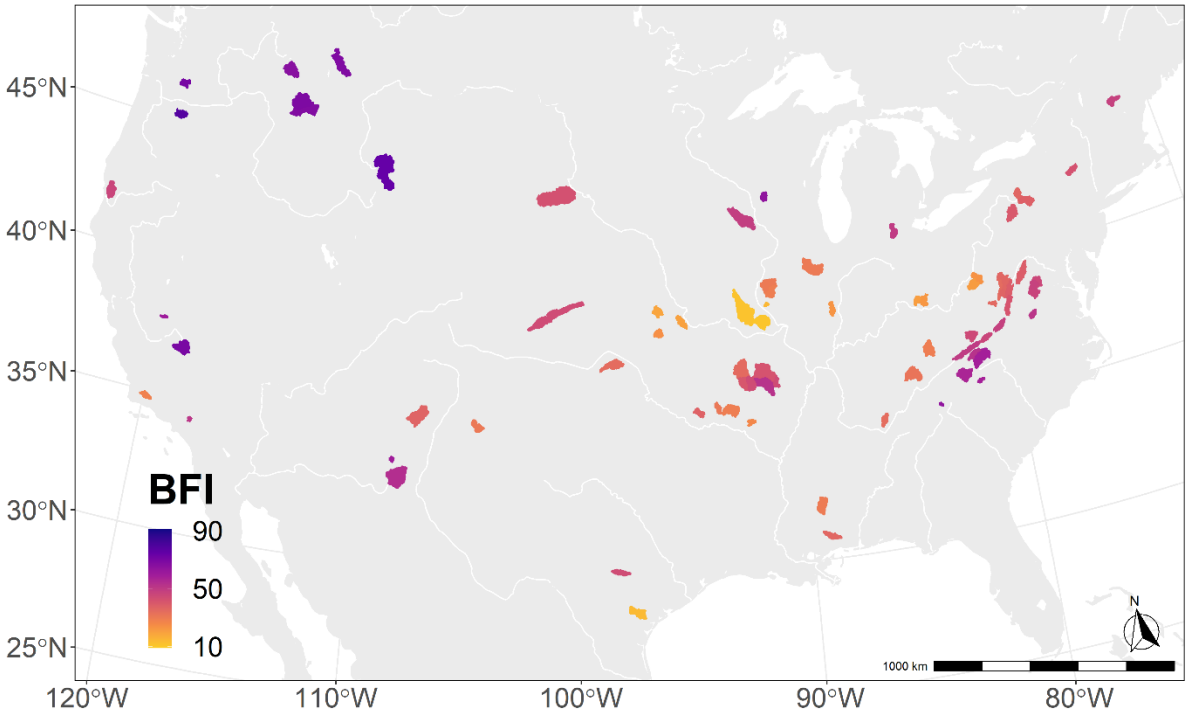

Figure 3: Map of MOPEX watersheds used in this study. Watersheds are highlighted according to their baseflow index (BFI [-]) as calculated using an automated hydrograph separation program [Wahl 1995] and reported by Wolock [2003].

1. **Implementing GrUB in HBV**

We implemented the GrUB module in HBV directly, by replacing the lower bucket (SLZ, see Table 1) and the lower bucket storage constant (K2) with Eq. 13. Specifically:

- N_0 from Eq. 13 was calculated according to the recharge term (Perc) from the intermediate bucket (SUZ) to the SLZ in HBV.IANIGLA
- Q_{sb} from Eq. 13 replaced Q_{bf} in HBV.IANIGLA
- S from Eq. 13 replaced SLZ in HBV.IANIGLA and was calculated according to N_0 and Q_{sb}

The GrUB module was never calibrated in any watershed. To assess the effects of the overlying model parameterization on the performance of GrUB, we did, however, implement GrUB into two differently calibrated versions of HBV (Figure 4). First, we calibrated HBV using all 14 parameters from Table 1 then

replaced the HBV groundwater module with GrUB. We refer to this implementation of HBV with the GrUB module as "GrUB-ind" as HBV was calibrated *ind*ependently of GrUB. We consider GrUB-ind to be directly comparable to the type of implementation that may be used in LSMs. Second, we calibrated HBV around the GrUB module by replacing K2 and SLZ with GrUB before beginning the calibration routine. We refer to this implementation of HBV with the GrUB module as "GrUB-co" because HBV was calibrated in *co*ordination with GrUB. We consider GrUB-co to be comparable to the type of implementation that may be used in rainfall-runoff models where stream gage data are available. Results from each of these implementations are shown alongside native HBV results in all subsequent figures.

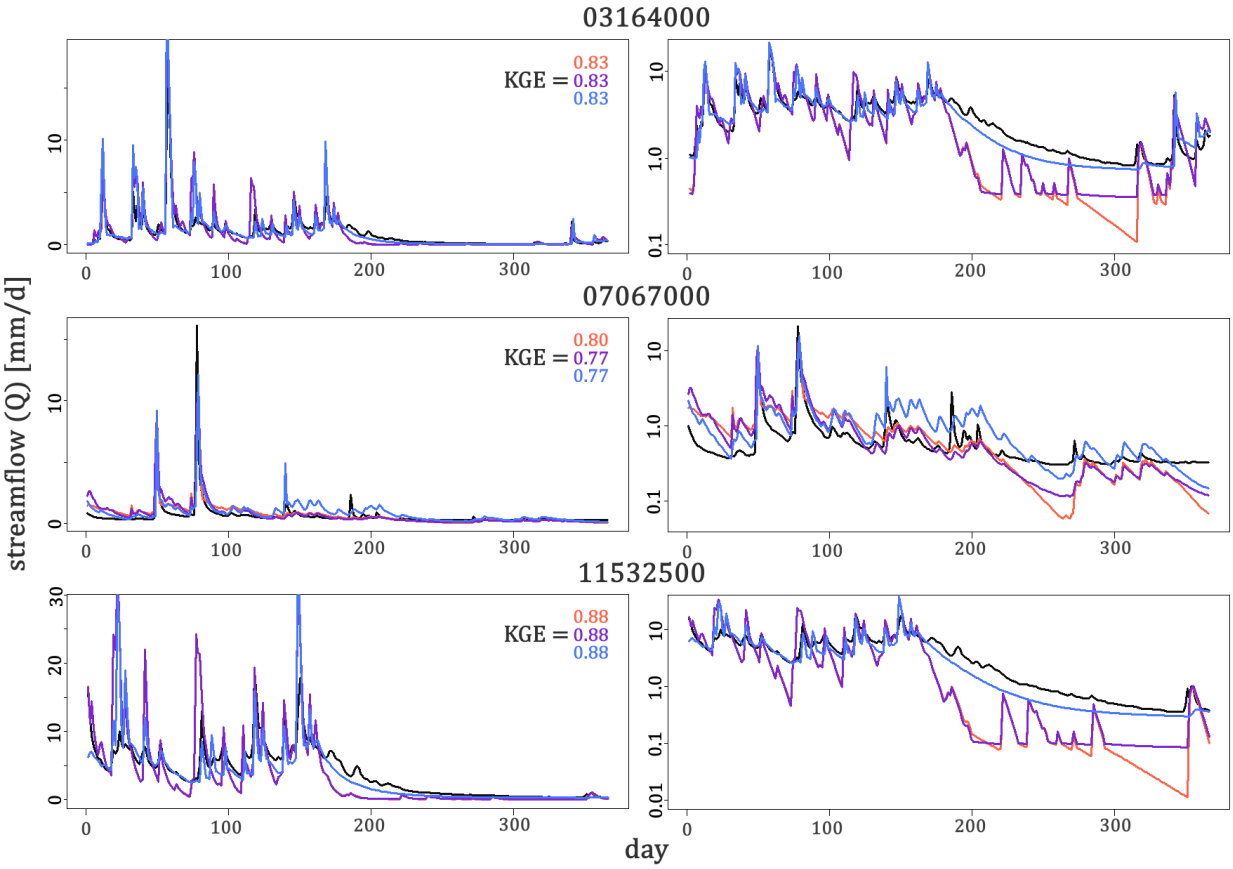

Figure 4: Example of model results for a single water year in three watershed: 1) USGS 03164000 New River near Galax, VA; 2) USGS 07067000 Current River at Van Buren, MO.; and 3) USGS 11532500 Smith River near Crescent City, CA. Gaged discharge is illustrated in black, HBV results in red, GrUB-ind in purple, and GrUB-co in blue. Plots on the right are in log space to highlight low flows.

1. Model Evaluation

HBV, GrUB-ind, and GrUB-co were evaluated using data from the entire simulation period according to their ability to accurately represent low flows. We identified periods of "low flow" according to two different methods. First, we applied the Tennant method [Tennant 1976], which is the most widely used method for defining minimum environmental instream flow in the US [Jowett 1997]. Minimum environmental instream flows are calculated as 30% of the mean annual flow for the period of record. Second, we applied a quantile threshold approach [Praskievicz et al. 2018]. Historical Q was ranked from largest to smallest, and streamflow records were selected based on an exceedance threshold of 96% (Q_{96}), which is often also referred to as the lower 4^{th} percentile of flow. The performance of each model during each of these subsets of low flows was then evaluated according to the mean absolute error (MAE), mean square error (MSE), and absolute bias.

1. **Model Results**

   (a) **General model performance**

The general evaluation results for all three models were similar (median KGE > 0.74) (Figure 5). HBV and GrUB-co generated nearly identical KGE results (1^{st} Q = 0.70, median= 0.75, 3^{rd} Q = 0.80, mean = 0.72). GrUB-ind marginally underperformed the other two models in most watersheds (1^{st} Q = 0.69, median= 0.74, 3^{rd} Q = 0.78, mean = 0.68). Distributions generated by GrUB-co and GrUB-ind were not statistically different from HBV ($p<0.05$) according to the two-sample Kolmogorov-Smirnov test [Smirnov 1948].

Overall performance of the three models was also similar according to mean absolute error (MAE), mean square error (MSE), absolute bias (Bias), and Nash-Sutcliffe efficiency (NSE) (Figure 5). GrUB-co generated the best results for all four metrics not used in the calibration routine (median MAE = 0.60 mm/day; median MSE = 2.02 mm/day; median Bias = 0.0003 mm/day; median NSE = 0.54). However, these represent statistically insignificant improvements over HBV values (median MAE = 0.61; median MSE = 2.03; median Bias = 0.006; and median NSE = 0.53). GrUB-ind generated the poorest results for three of the four performance metrics, though again these differences are not statistically significant (median MAE = 0.60; median MSE = 2.02; median Bias = 0.005; median NSE = 0.52).

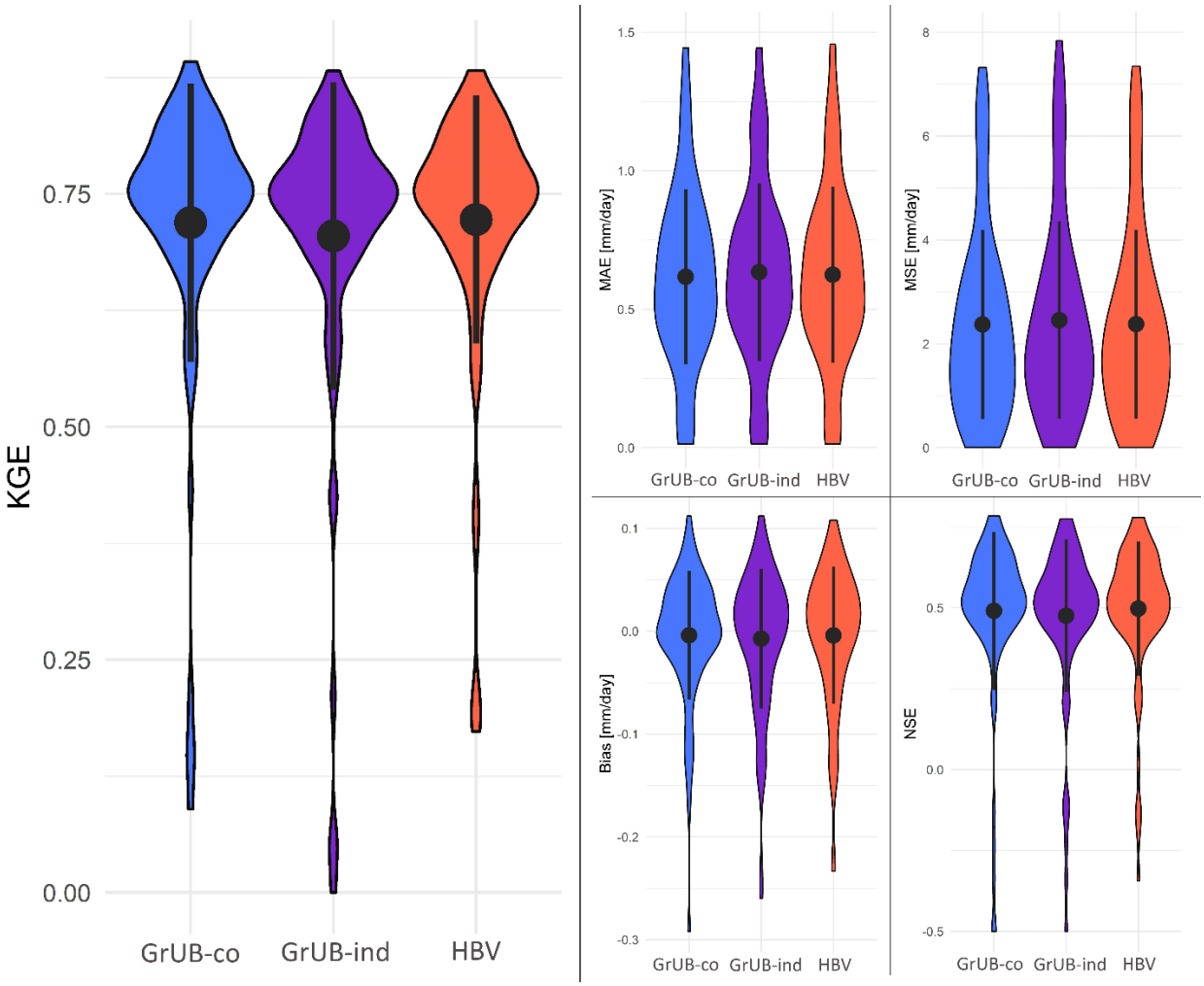

Figure 5: Violin plots of model evaluation results with KGE on left and MAE, MSE, Bias, and NSE on the right. Black dots represent mean values and black bars represent one standard deviation around the mean. Blue illustrates results from GrUB-co, purple from GrUB-ind, and red from HBV with the native calibrated groundwater module. Model performance results in the inset are (clockwise from top left) mean absolute error (MAE), mean square error (MSE), Nash-Sutcliffe efficiency (NSE), and absolute bias (Bias).

1. **Low flows**

All three models predicted low flows with similar accuracy, with GrUB-co and GrUB-ind generating slightly better results (Figure 6). For low flows identified according to Q_{96}, GrUB-co generated the lowest median values of MAE (0.10)

and MSE (0.022), while GrUB-ind generated the smallest median values of Bias (0.054). HBV generated the highest median values for each of these metrics, though only by a small margin (MAE = 0.11, MSE = 0.027, Bias = 0.066). All models tended to overpredict Q_{96} flows (67% of watersheds in GrUB-ind, 73% in GrUB=cal, and 69% in HBV). For low flows identified according to the Tennant method, GrUB-ind generated the lowest values of MAE (median = 0.17) and Bias (median = 0.09) while GrUB-co generated the lowest MSE (median = 0.12). HBV generated slightly higher values in all three categories (median MAE = 0.18, median MSE = 0.13, median Bias = 0.10). Distributions generated by GrUB-co and GrUB-ind were not statistically different from HBV ($p<0.05$) according to the two-sample Kolmogorov-Smirnov test.

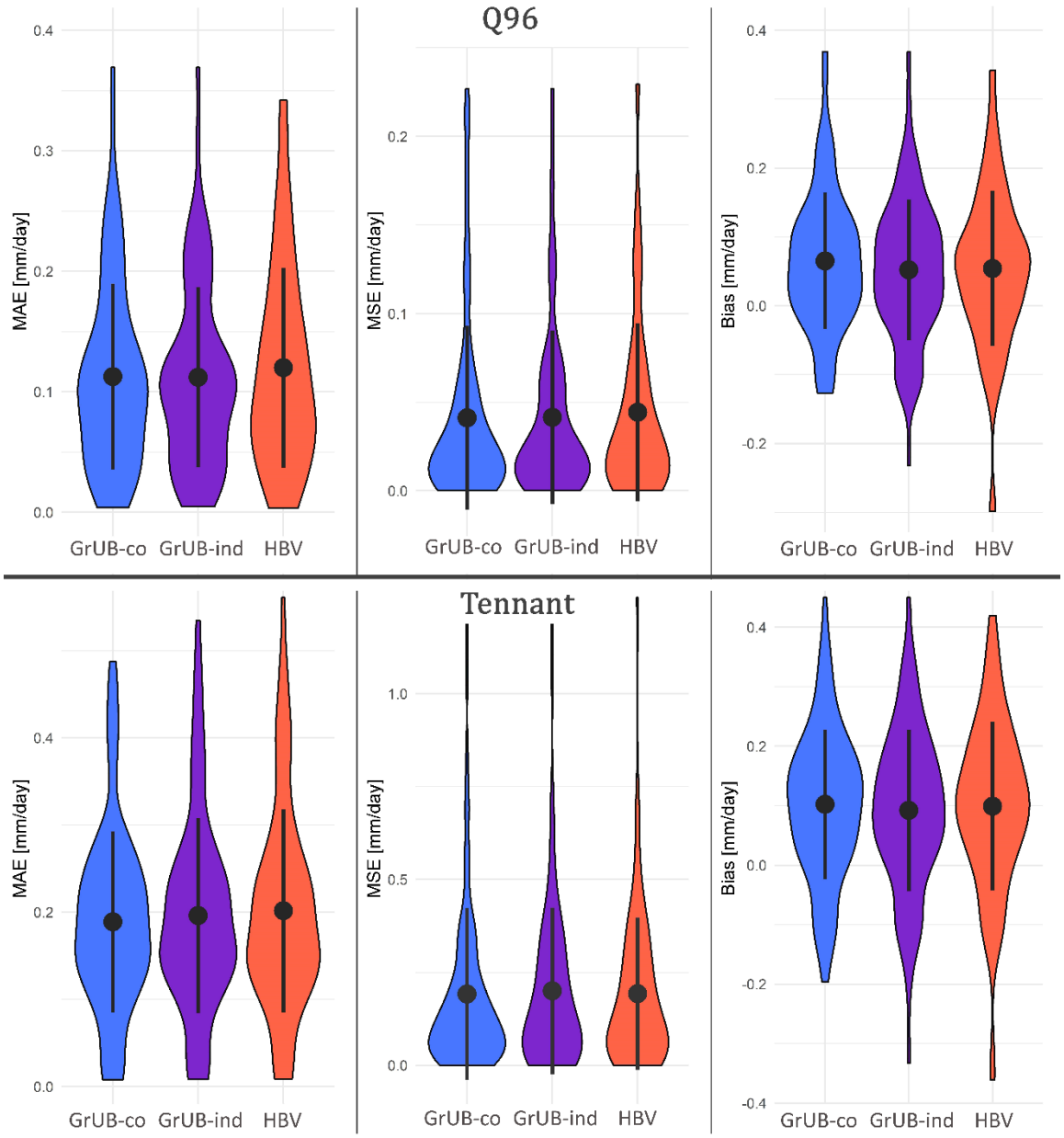

Figure 6: Violin plots of model predictions of low flows as defined as 96% exceedance (Q_{96}) flows (top row) and the Tennant method (bottom row). Blue illustrates results from GrUB-co, purple from GrUB-ind, and red from HBV with the native calibrated groundwater module. Model performance results are mean absolute error (MAE – left), mean square error (MSE - middle), and absolute bias

(Bias - right).

1. **Sensitivity to Rate of Recharge**

As our primary goal is to develop a groundwater module that is generally adaptable to a variety of LSMs and rainfall-runoff models, it is essential that we "stress test" our module to uncertainty in the structure of the overlying hydrologic model. Because recharge (N) is the only component of the overlying hydrologic model that interacts directly with GrUB, we chose to focus on components of the HBV model structure that constrain rates and magnitude of N. Specifically, we recalculated the rate at which infiltrated water percolates through the subsurface according to two end-member scenarios:

- Slow Recharge: UZL and Perc (see Table 1) set to maximum rates of 0.1 mm/day
- Fast Recharge: UZL and Perc set to minimum rates of 1000 mm/day

Recall (Table 1 and Figure 2) that UZL constrains the maximum rate of movement from soil storage to the upper storage zone (UZ) and Perc similarly constrains the maximum rate of movement from the UZ to the lower storage zone (LZ). Therefore, in the Slow Recharge scenario N is constrained to a small, nearly constant rate of 0.1 mm/day at maximum. Meanwhile, the Fast Recharge scenario effectively represents all water reaching the water table immediately following percolation below the shallow soils store. For all three models, we altered the values of UZL and Perc but otherwise kept all parameter values the same (i.e., the values listed in Table 1).

1. **Alternative Scenario: Slow Recharge**

As expected, all three models generated lower KGE and higher error metrics under the Slow Recharge scenario (Figure 7) than they did under the optimized scenario (Figure 5). GrUB-ind and HBV generated nearly identical overall error metrics (median values of KGE = 0.60; MSE = 3.6; NSE = 0.23), though HBV had higher median Bias (0.008 vs 0.004). GrUB-co generated superior or equivalent results in every category (median values of KGE = 0.63; MSE = 3.6; NSE = 0.30, Bias = 0.001). However, distributions generated by GrUB-co and GrUB-ind were not significantly different from HBV ($p<0.05$) according to the two-sample Kolmogorov-Smirnov test [Smirnov 1948].

This general pattern repeated itself when models were assessed during periods of low flow (Figure 8). GrUB-ind and HBV again generated nearly identical performance metrics under Q5 (median values of MAE = 0.11; MSE = 0.07; Bias = 0.021 and 0.022) while GrUB-co moderately outperformed each (median MAE = 0.10; MSE = 0.06; Bias = 0.021).

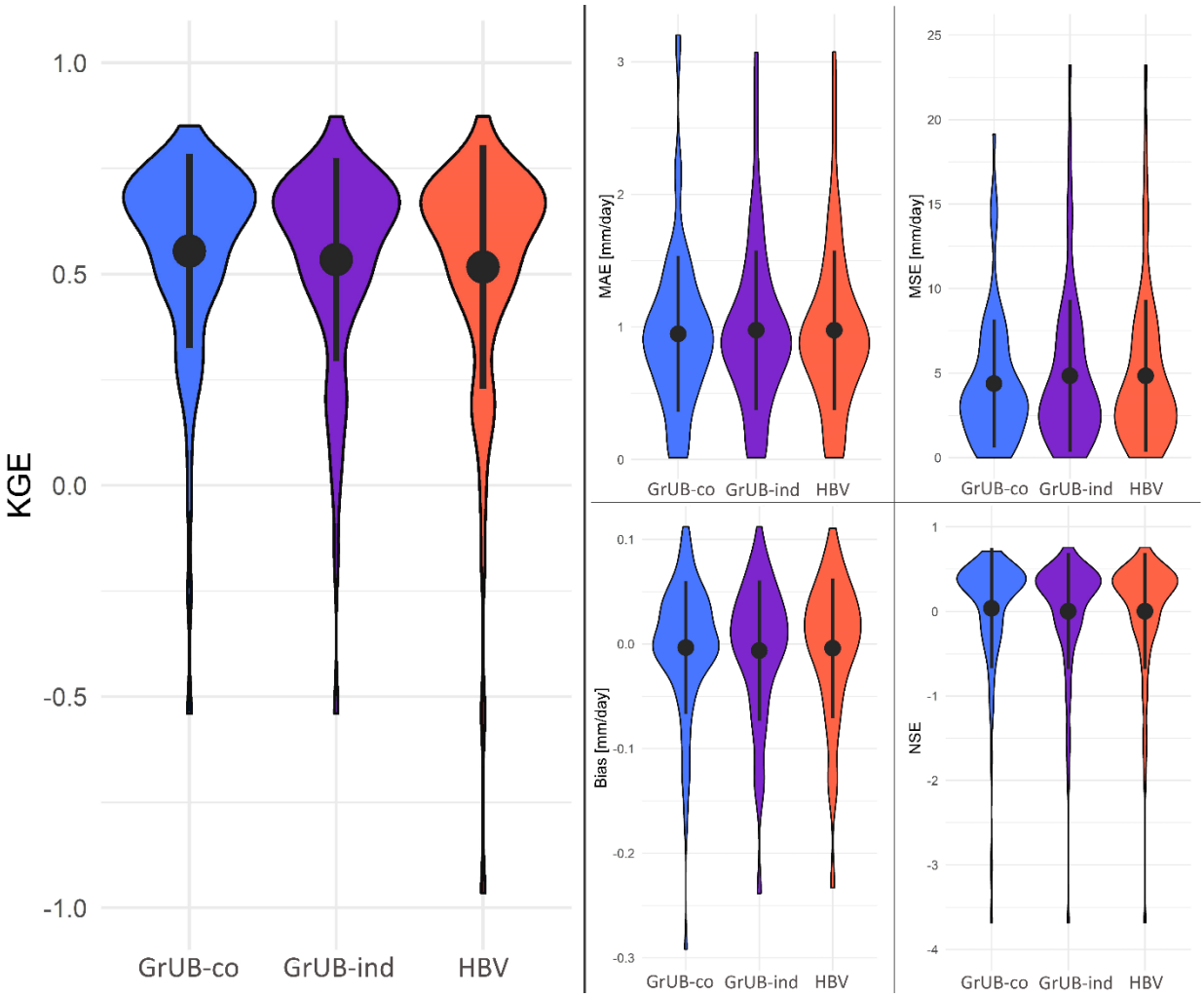

Figure 7: Violin plots of model performance under the alternative scenario of Slow Recharge. Black dots represent mean values and black bars represent one standard deviation around the mean. Blue illustrates results from GrUB-co, purple from GrUB-ind, and red from HBV with the native calibrated groundwater module. KGE results are on the left. Model performance results in the inset are (clockwise from top left) mean absolute error (MAE), mean square error (MSE), Nash-Sutcliffe efficiency (NSE), and absolute bias (Bias).

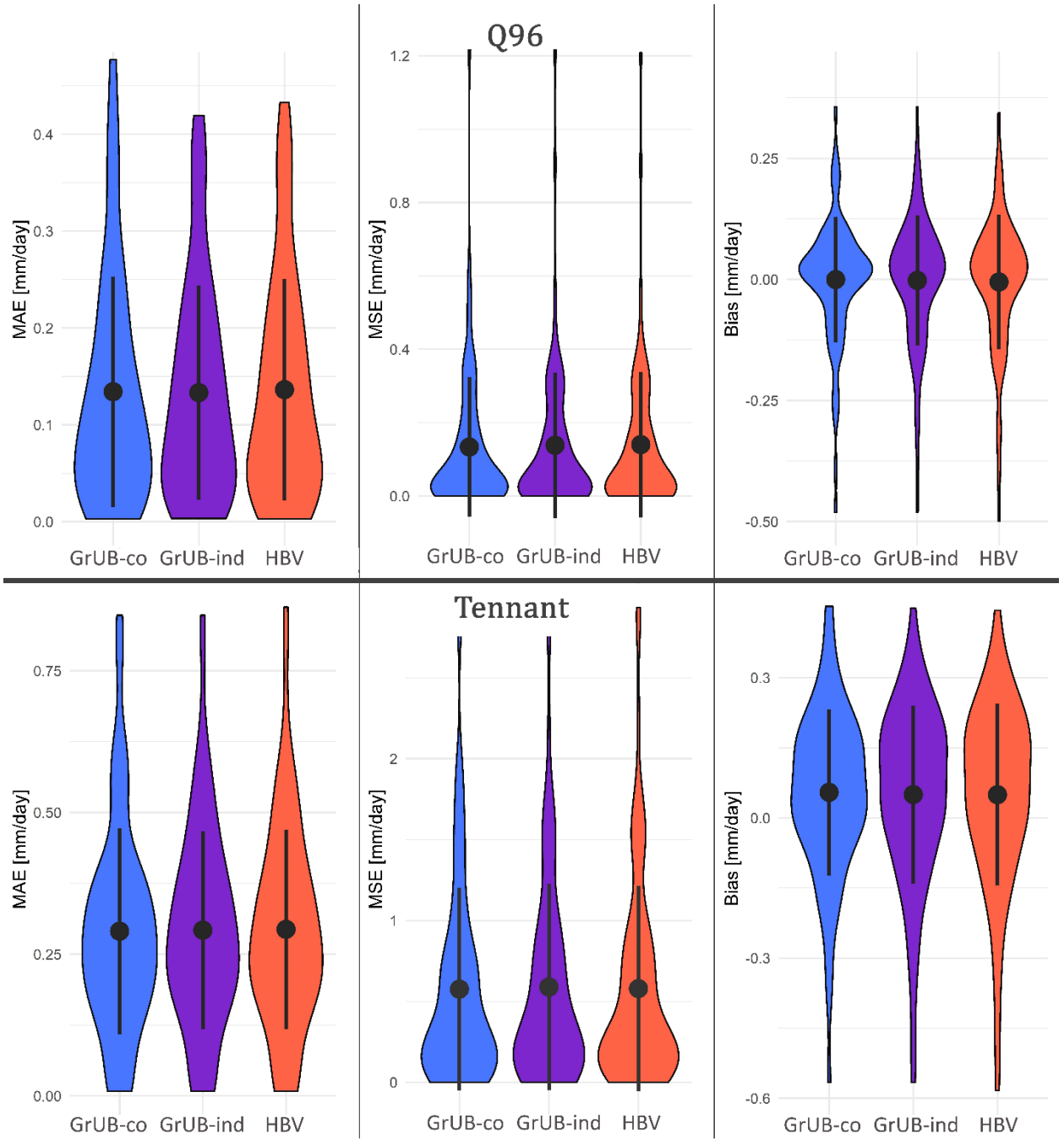

Figure 8: Violin plots of model predictions of low flows under the

alternative scenario of Slow Recharge. Low flows are defined as 96% exceedance (Q_{96}) flows (top row) and according to the Tennant method (bottom row). Blue illustrates results from GrUB-co, purple from GrUB-ind, and red from HBV with the native calibrated groundwater module. Model performance results are mean absolute error (MAE – left), mean square error (MSE - middle), and absolute bias (Bias - right).

1. **Alternative Scenario: Fast Recharge**

As expected, overall performance of all three models declined under the scenario of Fast Recharge. For the two metrics that preferentially weigh high flows (NSE and MSE), HBV outperformed both GrUB-co and GrUB-ind (median NSE = 0.1 vs -0.03; MSE = 3.5 vs 4.1) with statistical significance ($p<0.05$) according to the two-sample Kolmogorov-Smirnov test. However, GrUB-co and GrUB-ind each outperformed HBV by an equally wide margin on the other three metrics (median KGE = .08 vs 0.28; MAE = 0.85 vs 0.77; Bias = 0.009 vs 0.000), although only KGE distributions improved with statistical significance ($p<0.05$).

During periods of low flow, however, both GrUB-ind and GrUB-co generated nearly identical results that outperformed HBV by as much as an order of magnitude in every category. This held true for low flows identified by Q_{96} (median values of MAE = 0.26 vs 0.12; MSE = 0.10 vs 0.02; Bias = 0.26 vs 0.00) as well as low flows identified by the Tennant method (median values of MAE = 0.44 vs 0.20; MSE = 0.30 vs 0.10; Bias = 0.42 vs 0.17). All distributions generated by GrUB-co and GrUB-ind were significantly different from HBV ($p<0.05$) according to the two-sample Kolmogorov-Smirnov test [Smirnov 1948].

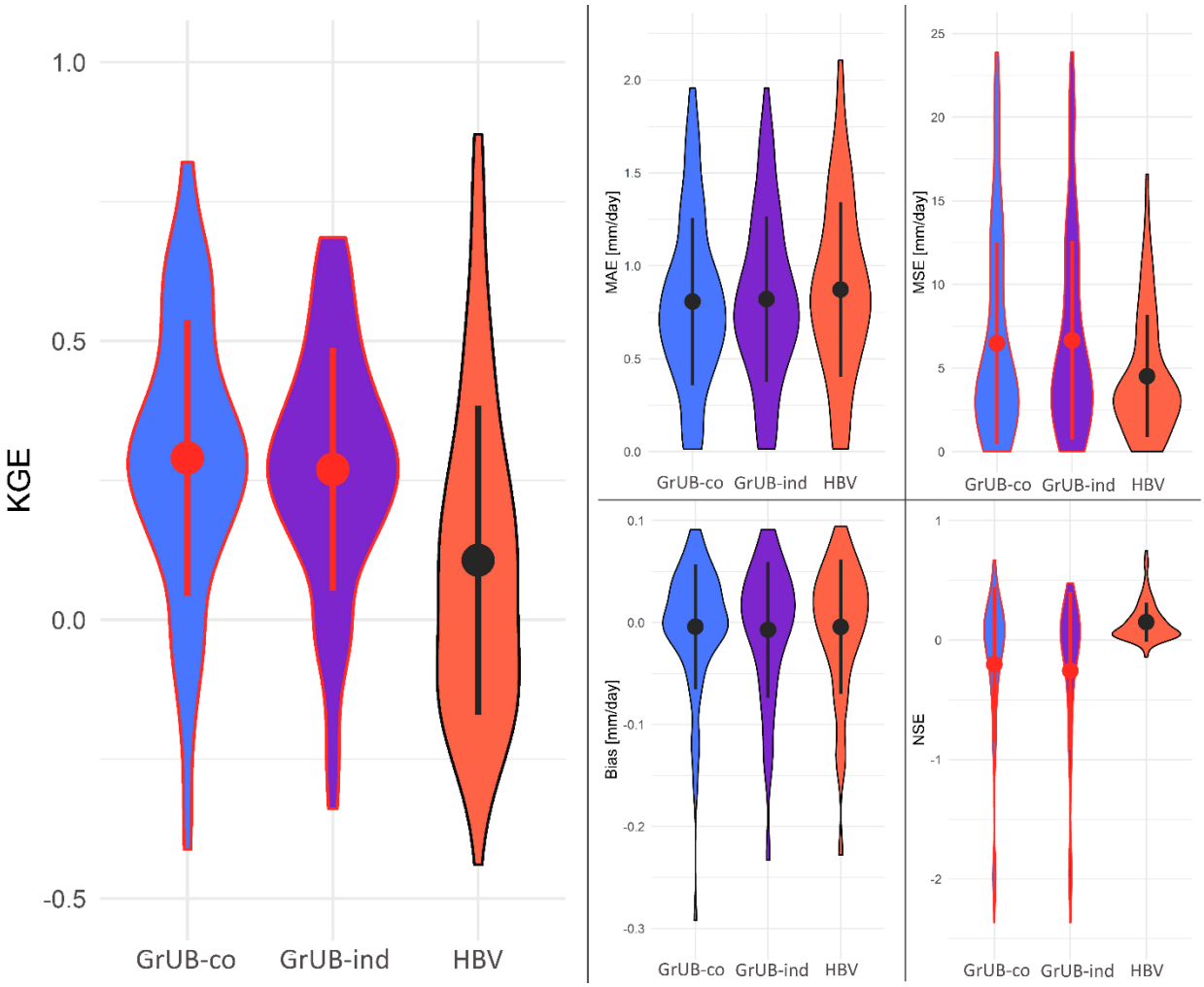

Figure 9: Violin plots of model performance under the alternative scenario of Fast Recharge. Black dots represent mean values and black bars represent one standard deviation around the mean. Blue illustrates results from GrUB-co, purple from GrUB-ind, and red from HBV with the native calibrated groundwater module. KGE results are on the left. Model performance results in the inset are (clockwise from top left) mean absolute error (MAE), mean square error (MSE), Nash-Sutcliffe efficiency (NSE), and absolute bias (Bias). Distributions that are significantly different from those of the basic HBV model are highlighted in red.

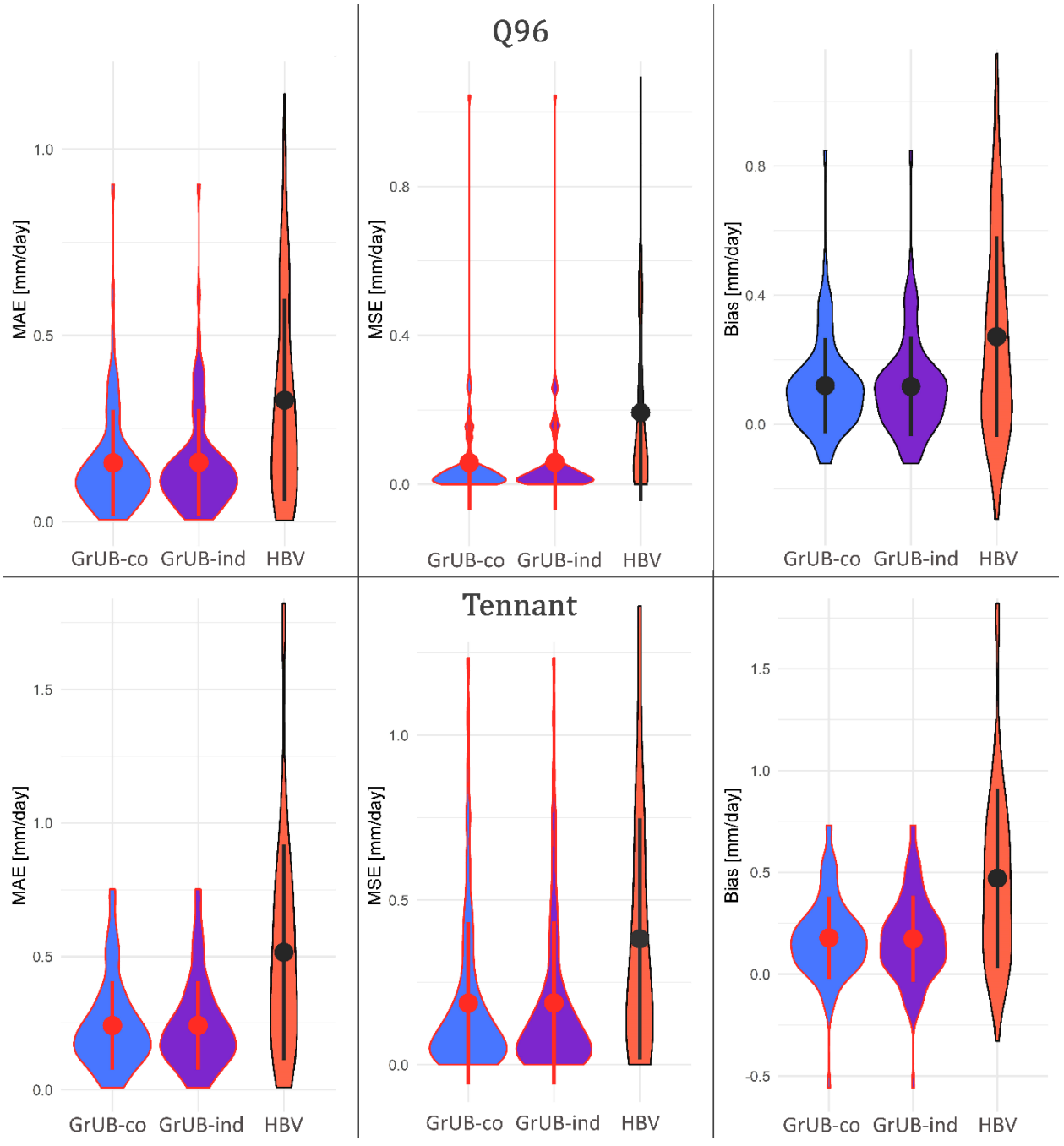

Figure 10: Violin plots of model predictions of low flows under the alternative scenario of Fast Recharge. Low flows are defined as

96% exceedance (Q_{96}) flows (top row) and according to the Tennant method (bottom row). Blue illustrates results from GrUB-co, purple from GrUB-ind, and red from HBV with the native calibrated groundwater module. Model performance results are mean absolute error (MAE – left), mean square error (MSE - middle), and absolute bias (Bias - right). Distributions that are significantly different from those of the basic HBV model are highlighted in red.

1. **Sensitivity to Model Parameterization**

GrUB relies on three parameters (p, v, and w) that are not derived from empirical data, but instead estimated as universal constants. We tested sensitivity to these parameters by re-running GrUB-ind for all watersheds with a wide range of parameter values. The range of parameter values was chosen to extend to (or beyond) values that are physically realistic.

For p, which represents the ratio of H_B to $H_{minimum}$, we selected values of 2, 10, 100, 200, and 1000. Because Th is the maximum value of H_B and median Th for the coterminous US is estimated at 3m [Tashie et al. 2021], a value of $p = 2$ represents a typical $H_{minimum}$ of 1.5 meters and a value of $p = 1000$ represents a typical $H_{minimum}$ of 3 mm. For v, which constrains the length of time that recharge (N) has a non-negligible effect on H_0 ($v^{-t} > 0.01$), we chose values of 0.1, 0.52, 0.95, 0.963, and 0.9875. These represent time periods of 2 days, 1 week, 1 season, 2 seasons, and 1 year. For w, which represents the exponential decay of $X_{h \sim H_B}$ as a function of S, we chose values of 1, 2, 3, 5, and 10. With w=1 the hillslope aquifer approaches a unit gradient when the aquifer is 99.7% saturated ($S = .997\ S_{max}$), and with w=10 it approaches a unit gradient when the aquifer is 48.0% saturated ($S = 0.480\ S_{max}$).

Parameter values did have noticeable effects on the hydrographs of individual watersheds, though these differences did not express themselves as significant changes in the overall performance of GrUB across all watersheds (Figure 11). We assessed the significance ($p<0.05$) of their effects on KGE (for the entire period of record) and MAE (during periods of low flow) using the Wilcoxon rank sum test [Wilcoxon 1945] and the two-sample Kolmogorov-Smirnov test [Smirnov 1948]. None were significant, with the lowest p values ($p = 0.35$) registering for changes in the variable p.

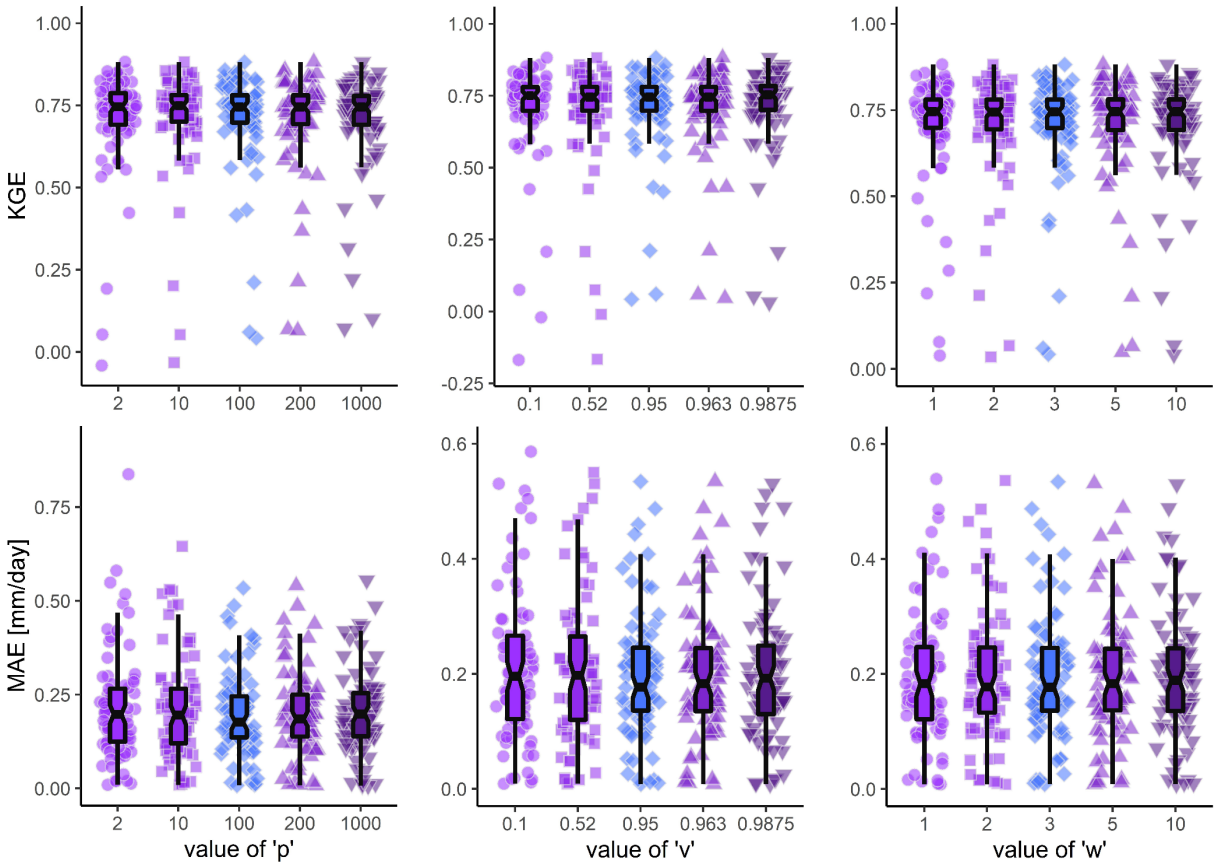

Figure 11: Boxplots of GrUB sensitivity to parameter selection. The top row illustrates KGE for the entire period of record. The bottom row illustrates mean absolute error (MAE) during low flows as defined by the Tennant method. The parameterization used in the primary analysis is illustrated with blue diamonds, with each diamond representing the value for a single watershed. Alternative parameter values are illustrated in purple circles, squares, and triangles.

1. **Model Performance**

We assess the overall performance of the GrUB module not in terms of its absolute performance in maximizing objective function (i.e., KGE) as this feat is largely achieved by the overlying hydrologic model (i.e., HBV). Instead, we assess its capacity to achieve the four key objectives (a-d) and three key performance metrics (i-iii) outlined in the introduction. The GrUB module does largely achieve its four key objectives:

1. No calibration required: the wholly uncalibrated GrUB module (GrUB-ind) is fully compatible with the otherwise calibrated HBV model.

2. Simple data requirements: parameterization of GrUB only requires data from several freely available continental- and global-scale datasets and requires minimal processing on the part of the model user. Unfortunately, three of the required parameters (K_0, m, and S_{max}) are, to our knowledge, currently only available for the coterminous US. Until these data are developed globally, the practical application of GrUB may be limited to this region.

3. Modularity: GrUB requires input from only a single, common flux term from the driving hydrologic model (i.e., deep recharge) and otherwise operates independently from the rest of the model structure.

4. Computational simplicity: though GrUB (Eq. 13) is more complex than the simple storage-discharge used in HBV, this added complexity is negligible in terms of the added computation time.

GrUB also largely achieves the three performance metric objectives, with some minor exceptions as outlined below.

1. Overall model performance is largely unaffected by incorporating either GrUB-ind or GrUB-co into HBV.

2. GrUB-co moderately improves predictions of low flows while GrUB-ind has a negligible effect.

3. Both GrUB-ind and GrUB-co are far more robust to changes in the overlying hydrologic model parameterization than is the standard HBV storage-discharge module.

For the potential incorporation of GrUB into LSMs, these results are generally promising, though with some important caveats (see Section 6). That the wholly uncalibrated GrUB-ind predicts low flows as well as a (simple) calibrated reservoir may prove beneficial to LSMs, which are broadly seen as describing low flows poorly without post-hoc calibration [Fan et al. 2019, Holtzman et al. 2020]. More important than the general performance, however, is the consistency with which a groundwater module predicts low flows when that module is forced by uncertain inputs from the overlying model structure. This "robustness" to changes in inputs from the overlying model is essential as different LSMs are structured and parameterized quite differently [Clark et al. 2015].

GrUB-ind is indeed extremely consistent in it prediction of low flows, as illustrated in the artificially extreme Slow Recharge and Fast Recharge scenarios (Section 4.3). Recall that the Slow Recharge scenario represents percolation rates of about .1 mm/day, compared with up to 1000 mm/day in the Fast Recharge scenario. Despite these potential rates of recharge spanning four orders of magnitude, error in GrUB-ind predictions of low flows was nearly identical to errors generated under calibrated conditions (median MAE during Slow, Fast, and calibrated conditions was 0.11, 0.13, and 0.11) (Figures 6, 8, and 10). Conversely, HBV error during low flows more than doubled under conditions of

Fast Recharge (median MAE during Slow, Fast, and calibrated conditions was 0.11, 0.28, and 0.11).

Presented results have mixed implications for the potential incorporation of GrUB into rainfall-runoff models. The hypothesis that reducing the number of variables that require calibration may allow for the superior optimization of the remaining parameters [e.g., Fraikin et al. 2019] is unfounded in this case, as the performance of GrUB-co is practically equivalent to that of HBV. Therefore, we do not expect GrUB to enhance the performance of other calibrated rainfall-runoff models where sufficient data for calibration is available. However, rainfall-runoff models are commonly applied in data sparse regions with "regionalization" approaches replacing parameter tuning via calibration (e.g., Swain and Patra 2017). Regionalization approaches tend to be subject to high uncertainty, with no single regionalization method yet accepted as a standard [Samuel et al. 2011, Oudin et al. 2008]. Because GrUB-ind is robust to changes in the parameterization of the overlying hydrologic model, it may prove useful in constraining model characterization of low flows in basins where regionalization approaches are necessary.

1. **Limitations and Future Efforts**

GrUB depends on three parameter values that are not currently available globally (K_0, m, and S_{max}). Though the methods to estimate these variables globally have been established [Tashie et al. 2021], until those data are available GrUB may not be applied outside the coterminous US. GrUB also relies on L, estimates of which are known to be extremely uncertain and which is known to vary with antecedent conditions by a factor of 2 or more [Godsey and Kirchner 2014]. While incorporating improved estimates of average L [Lin et al. 2021] may improve model performance, no large-scale data sets yet exist for describing its dynamic response to catchment conditions.

GrUB further relies on three parameters (p, v, and w) that are treated as empirical constants. In many hydrologic models, these parameters might present an opportunity for fine tuning model behavior. However, fine-tuning and unique calibrations are antithetical to the purpose of GrUB. Even though GrUB is not overly sensitive to changes in the value of these parameters (Section 4.4), the fact that these parameters are estimated according to hydrologic "intuition" rather than directly calculated represents an unfortunate deviation from the otherwise empirically-based structure of GrUB.

Though GrUB does incorporate seasonal-scale "watershed memory" by calculating H_0 as a function of recent recharge (Eq. 12), it does not account for potential multiannual long-term memory of climatic forcings that have been documented in many watersheds [e.g., Fowler et al. 2019]. Climate change is expected to induce long-term changes in watershed storage (and related ecohydrological responses) that are not directly accounted for in current generation LSMs [Argus et al. 2017, Enzminger et al. 2019]. In its current form, GrUB does not make significant strides towards addressing this deficiency. Current generation LSMs

are also known to underestimate evapotranspiration during dry periods. A potential explanation for this is that the representation of watersheds according to a representative soil column fails to account for the heterogeneity in soil moisture and depth to water table that is induced in actual watersheds by topography, geology, and soil structure. Though a hillslope hydraulics model of groundwater has been invoked as a potential tool for better capturing evapotranspiration during dry periods [Clark et al. 2015, Fan et al. 2019], we do not attempt to model such behavior here.

1. **Conclusion**

We develop a calibration-free, computationally simple module called Groundwater for Ungauged Basins (GrUB) to predict groundwater flow contribution to streamflow. GrUB may be readily incorporated into a variety of rainfall-runoff models and land surface models (LSMs). It requires no calibration, but instead depends entirely on empirical data that is available for the entire coterminous US and could potentially be derived globally. We assess the performance of GrUB in over 80 US watersheds by incorporating it into HBV, a popular rainfall-runoff model, and comparing overall performance metrics as well as error in predictions of low flows by the native (calibrated) HBV groundwater module and those by the GrUB module. The uncalibrated GrUB module generates error metrics that are equivalent to (or superior to) those generated by the calibrated HBV groundwater module. To ensure that predictions by GrUB are robust to changes in the structure and parameterization of the overlying hydrologic model, we run tests according to two artificial scenarios: Slow Recharge at a rate of up to 0.1 mm/day, and Fast Recharge at a rate of up to 1000 mm/day. GrUB proves to be very robust to these extreme changes, with mean absolute error (MAE) of predictions of low flows only increasing by an average of 3% and 19%, respectively, in the Fast and Slow recharge scenarios. We suggest GrUB as a potential tool for improving predictions of low flows in LSMs as well as rainfall-runoff models where calibration data are sparse.

Acknowledgements and Data:

The authors have no real or perceived conflicts of interest relating to this work. This research was funded by NSF grands EAR-1920425 and OIA-2019561. All data used in this study are publicly available on: United States Geological Survey (USGS) National Water Information System at https://waterdata.usgs.gov/nwis/rt (for stream discharge); CUAHSI's HydroShare platform at https://www.hydroshare.org/resource/115409dbe8354e78a2c2219d32e2b9de (for hydraulic properties) and at https://www.hydroshare.org/resource/99d5c1a238134ea6b8b767a65f440cb7 (for MOPEX data); and the USGS National Hydrography repository at https://www.usgs.gov/core-science-systems/ngp/national-hydrography/nhdplus-high-resolution (for catchment characteristics) (for catchment characteristics). All code used to read, analyze, and plot data is publicly available under DOI: 10.5281/zenodo.5019561 at https://zenodo.org/record/5019561#.YNNpYUwpBhE.

Works Cited

Allen, G. H., Pavelsky, T. M., Barefoot, E. A., Lamb, M. P., Butman, D., Tashie, A., & Gleason, C. J. (2018). Similarity of stream width distributions across headwater systems. Nature communications, 9(1), 1-7.

Argus, D. F., Landerer, F. W., Wiese, D. N., Martens, H. R., Fu, Y., Famiglietti, J. S., ... & Watkins, M. M. (2017). Sustained water loss in California's mountain ranges during severe drought from 2012 to 2015 inferred from GPS. Journal of Geophysical Research: Solid Earth, 122(12), 10-559.

Aulenbach, B. T., Hooper, R. P., van Meerveld, H. J., Burns, D. A., Freer, J. E., Shanley, J. B., ... & Peters, N. E. (2021). The Evolving Perceptual Model of Streamflow Generation at the Panola Mountain Research Watershed. Hydrological Processes, e14127.https://onlinelibrary.wiley.com/doi/10.1002/hyp.14127

Baker, I. T., Prihodko, L., Denning, A. S., Goulden, M., Miller, S., & da Rocha, H. R. (2008). Seasonal drought stress in the Amazon: Reconciling models and observations. Journal of Geophysical Research, 113, G00B01.

Balkovič, J., Skalský, R., Folberth, C., Khabarov, N., Schmid, E., Madaras, M., et al. (2018). Impacts and uncertainties of +2 °C of climate change and soil degradation on European crop cal orie supply. Earth's Future, 6(3), 373–395.

Batelis, S. C., Rahman, M., Kollet, S., Woods, R., & Rosolem, R. (2020). Towards the representation of groundwater in the Joint UK Land Environment Simulator. Hydrological Processes.

Bart, R., & Hope, A. (2014). Inter-seasonal variability in baseflow recession rates: The role of aquifer antecedent storage in central California watersheds. Journal of Hydrology, 519, 205-213.

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

Bergstrom, S. (1976) Development and Application of a Conceptual Runoff Model for Scandinavian Catchments. Report RH07, Swedish Meteorological and Hydrological Institute, Norrkoping.

Bergstrom, S. (1992) *The HBVmodel - its structure and applications.* Report. RH 4,Swedish Meteorological and Hydrological Institute, Norrkoping.

Bergström, S., & Lindström, G. (2015). Interpretation of runoff processes in hydrological modelling—experience from the HBV approach. Hydrological Processes, 29(16), 3535-3545.

Beven, K. (2006). A manifesto for the equi nality thesis. Journal of Hydrology, 320(1 2), 18–36.

Boughton, W. (2006). Calibrations of a daily rainfall runoff model with poor quality data. Environmental Modelling and Software, 21(8), 1114–1128.

Brunke, M. A., Broxton, P., Pelletier, J., Gochis, D., Hazenberg, P., Lawrence, D. M., Leung, L. R., Niu, G., Troch, P. A., & Zeng, X. (2016). Implementing and evaluating variable soil thickness in the Community Land Model, version 4.5 (CLM4.5). Journal of Climate, 29, 3441–3461

Brunner, M. I., Slater, L., Tallaksen, L. M., & Clark, M. (2021). Challenges in modeling and predicting floods and droughts: A review. Wiley Interdisciplinary Reviews: Water, e1520.

Carrer, G. E., Klaus, J., & Pfister, L. (2019). Assessing the catchment storage function through a dual-storage concept. Water Resources Research, 55(1), 476-494.Chegwidden, O. S., Rupp, D. E., & Nijssen, B. (2020). Climate change alters flood magnitudes and mechanisms in climatically-diverse headwaters across the northwestern United States. Environmental Research Letters, 15(9), 094048.

Clark, M. P., Rupp, D. E., Woods, R. A., Tromp-van Meerveld, H. J., Peters, N. E., & Freer, J. E. (2009). Consistency between hydrological models and field observations: linking processes at the hillslope scale to hydrological responses at the watershed scale. Hydrological Processes: An International Journal, 23(2), 311-319.

Clark, M. P., Fan, Y., Lawrence, D. M., Adam, J. C., Bolster, D., Gochis, D. J., ... & Zeng, X. (2015). Improving the representation of hydrologic processes in Earth System Models. Water Resources Research, 51(8), 5929-5956.

Cui, T., Yang, T., Xu, C. Y., Shao, Q., Wang, X., & Li, Z. (2018). Assessment of the impact of climate change on ow regime at multiple temporal scales and potential ecological implications in an alpine river. Stochastic Environmental Research and Risk Assessment,1–18

Dai, Y., Shangguan, W., Wei, N., Xin, Q., Yuan, H., Zhang, S., ... & Yan, F. (2019). A review of the global soil property maps for Earth system models. Soil, 5(2), 137-158.

Emanuel, R. E. (2018). Climate change in the Lumbee River watershed and potential impacts on the Lumbee tribe of North Carolina. Journal of Contemporary Water Research & Education, 163(1), 79–93.

Enzminger, T. L., Small, E. E., & Borsa, A. A. (2019). Subsurface water dominates Sierra Nevada seasonal hydrologic storage. Geophysical Research Letters, 46(21), 11993-12001.

Falcone, J. A., Carlisle, D. M., Wolock, D. M., & Meador, M. R. (2010). GAGES: A stream gage database for evaluating natural and altered flow conditions in the conterminous United States: Ecological archives E091-045. Ecology, 91(2), 621-621.

Fan, Y., Clark, M., Lawrence, D. M., Swenson, S., Band, L. E., Brantley, S. L.,

... & Kirchner, J. W. (2019). Hillslope hydrology in global change research and Earth system modeling. Water Resources Research, 55(2), 1737-1772.

Fang, K., Ji, X., Shen, C., Ludwig, N., Godfrey, P., Mahjabin, T., & Doughty, C. (2019). Combining a land surface model with groundwater model calibration to assess the impacts of groundwater pumping in a mountainous desert basin. Advances in Water Resources, 130, 12-28.

Ficklin, D. L., Maxwell, J. T., Letsinger, S. L., & Gholizadeh, H. (2015). A climatic deconstruction of recent drought trends in the United States. Environmental Research Letters, 10(4), 044009.

Flörke, M., Schneider, C., & McDonald , R. I. (2018). Water competition between cities and agriculture driven by climate change and urban growth. Nature Sustainability, 1(1), 51.

Fraikin, N., Funk, K., Frey, M., & Gauterin, F. (2019). Dimensionality reduction and identification of valid parameter bounds for the efficient calibration of automated driving functions. Automotive and Engine Technology, 4(1), 75-91.

Freeze, R. A., and J. A. Cherry (1979), Groundwater, 604 pp., Prentice-Hall, Englewood Cliffs, N. J.

Fowler, K., Knoben, W., Peel, M., Peterson, T., Ryu, D., Saft, M., ... & Western, A. Many commonly used rainfall-runoff models lack long, slow dynamics: implications for runoff projections. Water Resources Research, e2019WR025286. https://agupubs.onlinelibrary.wiley.com/doi/epdf/10.1029/2019WR025286

Gan, Y., Liang, X. Z., Duan, Q., Chen, F., Li, J., & Zhang, Y. (2019). Assessment and Reduction of the Physical Parameterization Uncertainty for Noah-MP Land Surface Model. Water Resources Research.

Godsey, S. E., & Kirchner, J. W. (2014). Dynamic, discontinuous stream networks: hydrologically driven variations in active drainage density, flowing channels and stream order. Hydrological Processes, 28(23), 5791-5803.

Gupta, H. V., Kling, H., Yilmaz, K. K. and 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, doi:10.1016/j.jhydrol.2009.08.003.

Hare, D. K., Helton, A. M., Johnson, Z. C., Lane, J. W., & Briggs, M. A. (2021). Continental-scale analysis of shallow and deep groundwater contributions to streams. Nature Communications, 12(1), 1-10.

Hariri, R. H., Fredericks, E. M., & Bowers, K. M. (2019). Uncertainty in big data analytics: survey, opportunities, and challenges. Journal of Big Data, 6(1), 1-16.

Holtzman, N. M., Pavelsky, T. M., Cohen, J. S., Wrzesien, M. L., & Herman, J. D. (2020). Tailoring WRF and Noah-MP to improve process representation

of Sierra Nevada runoff: Diagnostic evaluation and applications. Journal of Advances in Modeling Earth Systems, 12(3), e2019MS001832.

Huning, L. S., & AghaKouchak, A. (2020). Global snow drought hot spots and characteristics. Proceedings of the National Academy of Sciences, 117(33), 19753-19759.

van den Hurk, B. J., Viterbo, P., Beljaars, A. C. M., & Betts, A. K. (2000). Offline validation of the ERA40 surface scheme.

Iqbal, M., Dahri, Z., Querner, E., Khan, A., & Hofstra, N. (2018). Impact of climate change on ood frequency and intensity in the Kabul River Basin. Geosciences, 8(4), 114.

Jensen, CK, KJ McGuire, and PS Prince. 2016. Headwater stream length dynamics across four physiographic provinces of the Appalachian Highlands. Water Resources Research. https://onlinelibrary.wiley.com/doi/pdf/10.1002/hyp.11259

Jowett, I. G. (1997). Instream flow methods: a comparison of approaches. Regulated Rivers: Research & Management: An International Journal Devoted to River Research and Management, 13(2), 115-127.

de Jong, P., Tanajura, C. A. S., Sánchez, A. S., Dargaville, R., Kiperstok, A., & Torres, E. A. (2018). Hydroelectric production from Brazil's São Francisco River could cease due to climate change and inter annual variability. Science of the Total Environment, 634, 1540–1553

Kowalczyk, E. A., Stevens, L., Law, R. M., Dix, M., Wang, Y. P., Harman, I. N., ... & Ziehn, T. (2013). The land surface model component of ACCESS: description and impact on the simulated surface climatology. Aust. Meteorol. Oceanogr. J, 63(1), 65-82.

Krinner, G., Viovy, N., de Noblet-Ducoudré, N., Ogée, J., Polcher, J., Friedlingstein, P., ... & Prentice, I. C. (2005). A dynamic global vegetation model for studies of the coupled atmosphere-biosphere system. Global Biogeochemical Cycles, 19(1).

Kuppel, S., Fan, Y., & Jobbagy, E. G. (2017). Seasonal hydrologic buffer on continents: Patterns, drivers and ecological benefits. Advances in Water Resources, 102, 178–187.

Lin, P., Pan, M., Wood, E. F., Yamazaki, D., & Allen, G. H. (2021). A new vector-based global river network dataset accounting for variable drainage density. Scientific data, 8(1), 1-9.

Mahmoud, S. H., & Gan, T. Y. (2018). Impact of anthropogenic climate change and human activities on environment and ecosystem services in arid regions. Science of the Total Environment, 633, 1329–1344.

Mayer, P.M., S.K. Reynolds, M.D. McCutchen, and T.J. Canfield. Riparian buffer width, vegetative cover, and nitrogen removal effectiveness: A review

of current science and regulations. EPA/600/R-05/118. Cincinnati, OH, U.S. Environmental Protection Agency, 2006.

Mendoza, G.F.; Steenhuis, T.S.; Walter, M.T.; Parlange, J.-Y. Estimating basin-wide hydraulic parameters of a semi-arid mountainous watershed by recession-flow analysis. J. Hydrol. 2003, 279, 57–69.

Milly, P. C., & Shmakin, A. B. (2002). Global modeling of land water and energy balances. Part I: The land dynamics (lad) model. Journal of Hydrometeorology, 3, 283–299.

Miguez-Macho, G., & Fan, Y. (2012a). The role of groundwater in the Amazon water cycle: 1. Influence on seasonal streamflow, flooding and wetlands. Journal of Geophysical Research, 117, D15113. https://doi.org/10.1029/2012JD017539

Miguez-Macho, G., & Fan, Y. (2012b). The role of groundwater in the Amazon water cycle: 2. Influence on seasonal soil moisture and evapotranspiration. Journal of Geophysical Research, 117, D15114. https://doi.org/10.1029/2012JD017540

Niu, G. Y., Yang, Z. L., Dickinson, R. E., & Gulden, L. E. (2005). A simple TOPMODEL-based runoff parameterization (SIMTOP) for use in global climate models. Journal of Geophysical Research: Atmospheres, 110(D21).

NOAA (2016), National Water Model: Improving NOAA's Water Prediction Services, http://water.noaa.gov/documents/wrn-national-water-model.pdf. 2 pp.

Oudin, L., Andréassian, V., Perrin, C., Michel, C., & Le Moine, N. (2008). Spatial proximity, physical similarity, regression and ungaged catchments: A comparison of regionalization approaches based on 913 French catchments. Water Resources Research, 44(3).

Padrón, R. S., Gudmundsson, L., Decharme, B., Ducharne, A., Lawrence, D. M., Mao, J., ... & Seneviratne, S. I. (2020). Observed changes in dry-season water availability attributed to human-induced climate change. Nature Geoscience, 13(7), 477-481.

Pelletier, J. D., Broxton, P. D., Hazenberg, P., Zeng, X., Troch, P. A., Niu, G. Y., ... & Gochis, D. (2016). A gridded global data set of soil, intact regolith, and sedimentary deposit thicknesses for regional and global land surface modeling. Journal of Advances in Modeling Earth Systems, 8(1), 41-65.

Pokhrel, Y. N., Fan, Y., Miguez-Macho, G., Yeh, P. J. F., & Han, S. C. (2013). The role of groundwater in the Amazon water cycle: 3. Influence on terrestrial water storage computations and comparison with GRACE. Journal of Geophysical Research: Atmospheres, 118, 3233– 3244. https://doi.org/10.1002/jgrd.50335

Praskievicz, S., Luo, C., Bearden, B., & Ernest, A. (2018). Evaluation of low-flow metrics as environmental instream flow standards during long-term average

and 2016 drought conditions: Tombigbee River Basin, Alabama and Mississippi, USA. Water Policy, 20(6), 1240-1255.

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rahman, M., Rosolem, R., Kollet, S. J., & Wagener, T. (2019). Towards a computationally efficient free-surface groundwater flow boundary condition for large-scale hydrological modelling. Advances in Water Resources, 123, 225–233.

Saft, M., Peel, M. C., Western, A. W., & Zhang, L. (2016). Predicting shifts in rainfall runoff partitioning during multiyear drought: Roles of dry period and catchment characteristics. Water Resources Research, 52(12), 9290 –9305.

Samuel, J., Coulibaly, P., & Metcalfe, R. A. (2011). Estimation of continuous streamflow in Ontario ungauged basins: comparison of regionalization methods. Journal of Hydrologic Engineering, 16(5), 447-459.

Sælthun, N. R. (1996). The"Nordic"HBV model. Description and documentation of the model version developed for the project Climate Change and Energy Production.

Schaake, J., Cong, S., & Duan, Q. (2006). US MOPEX data set (No. UCRL-JRNL-221228). Lawrence Livermore National Lab.(LLNL), Livermore, CA (United States).

Seibert, J., & van Meerveld, H. J. (2016). Hydrological change modeling: Challenges and opportunities. Hydrological Processes, 30(26), 4966–4971.

Shaw, S. B., & Riha, S. J. (2012). Examining individual recession events instead of a data cloud: Using a modified interpretation of dQ/dt–Q streamflow recession in glaciated watersheds to better inform models of low flow. Journal of hydrology, 434, 46-54.

Smakhtin, V. U. (2001). Low flow hydrology: a review. Journal of hydrology, 240(3-4), 147-186.

Smirnov, N (1948). Table for estimating the goodness of fit of empirical distributions. Annals of Mathematical Statistics.

Subin, Z., P. Milly, B. Sulman, S. Malyshev, and E. Shevliakova (2014), Resolving terrestrial ecosystem processes along a subgrid topographic gradient for an earth-system model, Hydrol. Earth Syst. Sci. Discuss., 11, 8443–8492.

Swain, J. B., & Patra, K. C. (2017). Streamflow estimation in ungauged catchments using regionalization techniques. Journal of Hydrology, 554, 420-433. Tague, C., and L. Band, 2004: Rhessys: Regional hydro-ecologic simulation system—an object-oriented approach to spatially distributed modeling of carbon, water, and nutrient cycling.Earthinteractions,8 (19), 1–42.

Tague, C., & Grant, G. E. (2009). Groundwater dynamics mediate low-flow response to global warming in snow-dominated alpine regions. Water resources research, 45(7).

Tashie, A., Scaife, C. I., & Band, L. E. (2019). Transpiration and subsurface controls of streamflow recession characteristics. Hydrological Processes, 33(19), 2561-2575.

Tashie, A., Pavelsky, T., & Band, L. E. (2020a). An empirical reevaluation of streamflow recession analysis at the continental scale. Water Resources Research, 56(1), e2019WR025448.

Tashie, A., Pavelsky, T., & Emanuel, R. E. (2020b). Spatial and temporal patterns in baseflow recession in the continental United States. Water Resources Research, 56(3), e2019WR026425.

Tashie, A., Pavelsky, T., Band, L., & Topp, S. (2021). Watershed-Scale Effective Hydraulic Properties of the Continental United States. Journal of Advances in Modeling Earth Systems, 13(6), e2020MS002440.

Tennant, D. L. (1976). Instream flow regimens for fish, wildlife, recreation and related environmental resources. Fisheries, 1(4), 6-10.

Toum E (2021). HBV.IANIGLA: Modular Hydrological Model. 0.2.0, https://CRAN.R-project.org/package=HBV.IANIGLA

Troch, P. A., Berne, A., Bogaart, P., Harman, C., Hilberts, A. G., Lyon, S. W., ... & Verhoest, N. E. (2013). The importance of hydraulic groundwater theory in catchment hydrology: The legacy of Wilfried Brutsaert and Jean-Yves Parlange. Water Resources Research, 49(9), 5099-5116.

USGS – United States Geological Survey. (2004). National Hydrography Dataset [Map]. [Reston, Va.] :U.S. Dept. of the Interior, U.S. Geological Survey

Wahl, K.L., and Wahl, T.L., 1995, Determining the flow of Comal Springs at New Braunfels, Texas, in Proceedings of Texas Water '95, August 16-17, 1995, San Antonio, Texas: American Society of Civil Engineers, p. 77-86, information available on the World Wide Web, accessed March 25, 2003, at URL http://www.usbr.gov/pmts/hydraulics_lab/twahl/bfi/texaswater95/comalsprings.html

Wieczorek, M. E., Jackson, S. E., & Schwarz, G. 2. (2018). Select attributes for NHDPlus Version 2.1 reach catchments and modified network routed upstream watersheds for the conterminous United States. Reston, VA: US Geological Survey.

Wilcoxon F (1945) Individual comparisons by ranking methods. Biometrics 1:80–83

Wolock, D.M., 2003a, Flow characteristics at U.S. Geological Survey streamgages in the conterminous United States: U.S. Geological Survey Open-File Re-

port 03-146, digital dataset, available on the World Wide Web, accessed June 30, 2003, at URL https://water.usgs.gov/lookup/getspatial?qsitesdd

Yang, Z.-L., Cai, X., Zhang, G., Tavakoly, A. A., Jin, Q., Meyer, L. H., & Guan, X. (2011). The community Noah land surface model with multi-parameterization options (Noah-MP). Austin, TX: The University of Texas at Austin.

Zambrano-Bigiarini, M. (2020). hydroGOF: Goodness-of-fit functions for comparison of simulated and observed hydrological time series. R package version 0.4-0. URL https://github.com/hzambran/hydroGOF. DOI:10.5281/zenodo.839854.

Zecharias, Y. B., & Brutsaert, W. (1988). Recession characteristics of groundwater outflow and base flow from mountainous watersheds. Water Resources Research, 24(10), 1651-1658.

Zhang, Y., & Schaap, M. G. (2019). Estimation of saturated hydraulic conductivity with pedotransfer functions: A review. Journal of Hydrology, 575, 1011–1030.

Zimmer, M. A., & McGlynn, B. L. (2017a). Ephemeral and intermittent runoff generation processes in a low relief, highly weathered catchment. Water Resources Research, 53(8), 7055-7077.

Zimmer, M. A., & McGlynn, B. L. (2017b). Time-lapse animation of hillslope groundwater dynamics details event-based and seasonal bidirectional stream–groundwater gradients. Hydrological Processes, 31(10), 1983-1985.