A Signature-based Hydrologic Efficiency Metric for Model Calibration and Evaluation in Gauged and Ungauged Catchments

Melike Kiraz[1*], Gemma Coxon[2], and Thorsten Wagener[1,3]

[1]University of Bristol, Civil Engineering, Bristol, United Kingdom

[2]University of Bristol, School of Geographical Sciences, Bristol, United Kingdom

[3]University of Potsdam, Institute for Environmental Science and Geography, Germany

*Corresponding author: Melike Kiraz (melike.kiraz@bristol.ac.uk)

Key Points:

- The underlying components of widely used efficiency metrics can be represented through different signatures
- These components can be estimated in ungauged basins and hence the metric itself can be calculated for ungauged model evaluation/calibration
- Modellers should replace the components of the proposed metric with signatures best suited to their research domains

Abstract

Rainfall-runoff models are commonly evaluated against statistical evaluation metrics. However, these metrics do not provide much insight into what is hydrologically wrong if a model fails to simulate observed streamflow well and they are also not applicable for ungauged catchments. Here, we propose a signature-based hydrologic efficiency (SHE) metric consisting of hydrologic signatures that can be regionalized for model evaluation in ungauged catchments. We test our new efficiency metric across 633 catchments from Great Britain. Strong correlations with Spearman rank and Pearson correlation values around 0.8 are found between our proposed metric and commonly used statistical evaluation metrics (NSE, KGE, NP…) demonstrating that the proposed SHE metric is related to existing metrics as much as these metrics are related to each other. For ungauged catchments, we regionalise the three signatures included in SHE and find that 78% of catchments have an absolute difference of SHE values between gauged and ungauged cases of less than 0.2. This difference increases where the regionalized bias and variance signature values are different to the observed ones. It means that SHE metric is applicable for model evaluation in ungauged catchments if its signatures can be regionalized well.

Keywords: Model evaluation, hydrologic signatures, evaluation metric, Great Britain, ungauged catchment

1 Introduction

Statistical objective functions are widely used to quantify the difference between observed and simulated streamflow time series for rainfall-runoff model evaluation and calibration in situations where historical streamflow observations are available. Such objective functions integrate the differences between observed and simulated time series, i.e. the residuals. Many metrics are based on the mean squared error (MSE) which can be derived from basic statistical assumptions about the errors present (Gershenfeld, 1999). In hydrology, Nash and Sutcliffe (1970) suggested that this metric should be normalized to allow for a better comparison of model performances across catchments. Their unit-free objective function has become well known as the Nash Sutcliffe Efficiency (NSE).

Multiple authors subsequently pointed out that metrics based on MSE type assumptions can be broken up into several constituent components, i.e. bias, standard deviation and correlation (Murphy, 1988; Weglarczyk, 1998). However, these components are not equally weighted within the traditional NSE formulation. Gupta et al. (2009) therefore suggested to combine them using Euclidean distance, which weights them equally in their Kling Gupta Efficiency (KGE) (see also Kling et al., 2012). This KGE metric has been used widely since its introduction and some authors have suggested improvements. For example, Pool et al. (2018) proposed to make the constituent components non-parametric so that they are less dependent on underlying assumptions. They replaced Pearson's linear correlation with Spearman rank correlation, and they assessed discharge variability using a normalized flow duration curve (FDC) to remove volume information and retain information about distributions only.

These metrics are undoubtedly cornerstones of hydrologic modelling, but some underlying problems with their use have been the basis for an ongoing debate. First, it is difficult to interpret them and their constituent components hydrologically (Gupta et al., 2008). For example, what is hydrologically wrong with my model if the NSE value is only 0.5? This problem has led to the use of hydrologic signatures in model evaluation (e.g. Moges et al., 2022).

Such signatures are indices of hydrologic function, such as the runoff ratio, which is an index that quantifies the fraction of precipitation that leaves the catchment as streamflow rather than evapotranspiration (McMillan, 2021). Second, the use of hard performance thresholds, though promoted by some (e.g. Moriasi et al., 2007; Rogelis et al., 2016; Towner et al., 2019), has been heavily criticized by others (e.g. Knoben et al., 2019; Clark et al., 2021). Flexible performance benchmarks have also been suggested to overcome this problem (e.g Seibert, 2001; Schaefli and Gupta, 2007; Seibert et al., 2018), while a more diagnostic evaluation of the underlying components has been proposed by others (Schwemmle et al., 2021).

Metrics like NSE and KGE are only applicable to gauged catchments because they require historical time series of observed streamflow to estimate residuals. However, previous studies have regionalized hydrologic signatures (e.g. Yadav et al., 2007; Hrachowitz et al., 2014; Pool and Seibert, 2021; Guo et al., 2021), and the statistical hydrology literature is rich with examples where streamflow statistics have been regionalized (e.g. Vogel et al., 1999). Therefore, at least some of the components that make up efficiency metrics, i.e., bias and variance, have already been estimated in ungauged basins. Indeed, there have been quite a few studies that have used (uncertain) regionalized hydrologic signatures as constraints for rainfall-runoff model ensembles (e.g. Zhang et al., 2008; Bulygina et al., 2009; Westerberg et al., 2011). However, there has been no attempt so far to build an efficiency metric for ungauged basins from these components.

In this paper, we propose a signature-based hydrologic efficiency metric that builds upon the work that has been done previously with signatures in both gauged and ungauged catchments. Integration of hydrologic signatures in an evaluation metric will provide opportunity for hydrologic interpretation of model performance and being able to regionalize these signatures will provide hydrologic efficiency evaluation of models for ungauged catchments. We test our ideas across 633 catchments in Great Britain (GB) by using model simulations in a Monte Carlo framework for a 10-year time period.

2 Data

In this paper, we analyse 633 catchments spread across Great Britain. Great Britain – consisting of England, Wales, and Scotland – is characterized by a temperate climate, moderate topographic variability, and significant geological heterogeneity. Precipitation decreases from north-west to south-east with a mean annual values ranging from 550 to 3500 mm/year (Coxon et al., 2020). Conversely, potential evapotranspiration (PET) increases from north-west (minimum of about 350 mm/year) to south-east (maximum of about 550 mm/day). Most of England is dominated by lowland terrain, whereas Wales and Scotland are dominated by more mountainous regions. Great Britain has a diverse geology including aquifers consisting of more permeable Chalk, Magnesian, Jurassic, Devonian/Carbonifero limestone and Permo-Triassic sandstone.

This study uses daily rainfall, streamflow, potential evapotranspiration time series for ten years (October 1, 1999 – September 30, 2009) and catchment attributes from the CAMELS-GB dataset to develop and demonstrate the new metric. CAMELS-GB is a large sample, open-source, hydro-meteorological dataset for Great Britain. It includes hydro-meteorological time series (consisting of rainfall, streamflow, potential evapotranspiration, temperature, radiation and humidity for 1970-2015 years), catchment attributes (including topography, climate, hydrology, land cover, soils, hydrogeology and human influences) (see Table S1) and catchment boundaries

for 671 catchments across Great Britain (Coxon et al., 2020). Considering climatic variability (i.e. wet and dry periods), ten-years of data is assumed to be sufficient to capture long-term climatic and hydrologic characteristics of our catchments for the purpose of this study. About 96% of the 671 catchments have >90% complete streamflow data in this 10-year period (i.e. 1999-2009). From the 671 CAMELS-GB catchments, we exclude 12 catchments from the analysis where (1) the runoff ratio or variance ratio value is higher than 1 – suggesting significant and unexplained water balance issues, (2) there is no available BFI-HOST data or (3) there is insufficient streamflow data for the specified study years. In addition, we also exclude 26 catchments where water balance analysis (see Section S3 in supplemental information) shows that they are significantly losing water most likely through subsurface processes which is not captured by the hydrological model used in this study. Hence, 633 GB catchments are used in the subsequent analysis.

3 Methodology

3.1 A Signature-based Hydrologic Efficiency (SHE) metric

We follow previous work discussed in the introduction section by adding a particular focus on signatures representing different hydrological dynamics as the individual components underlying hydrological efficiency metrics, as well as our ability to regionalize them (see Table 1).

a) Bias term: Runoff ratio

Runoff ratio (RR) is defined as the ratio of long-term average streamflow to long-term average precipitation. It is the long-term water balance separation between water being released from the catchment as streamflow and as evapotranspiration (Milly, 1994; Sankarasubramanian et al., 2001; Olden and Poff, 2003; Yadav, 2007). Higher runoff ratios identify catchments where a large amount of water leaves the catchment as streamflow with respect to precipitation and vice versa.

b) Variance (i.e. amplitude) term: Variance ratio

We define variance ratio as the ratio of standard deviation of streamflow to standard deviation of precipitation. The signature shows how variable (i.e. flashy) streamflow is with respect to precipitation drivers and is as such an indicator of the damping of precipitation variability through the catchment (a lower value indicating more damping).

c) Correlation term

Correlation is an aspect that is more difficult to capture in a signature. It could be represented as a function of the catchment response in relation to precipitation using the time of concentration of a catchment. However, estimates of time of concentration using the daily data we use in this study do not work very well for small and fast responding catchments in Great Britain (Giani et al., 2021). While exploration of this signature is beyond this technical note, we will return to the issue when we discussed ungauged basins. For now, we decided to use Spearman rank correlation between observed and simulated streamflow values as the correlation term of SHE like the non-parametric form of KGE developed by Pool et al. (2018). The components and formulation of SHE for gauged cases (i.e. SHE_g) are given in Table 1.

Table 1. Bias, variance and correlation components and formulations of evaluation metrics.

Objective function	Bias (β)	Variance (α)	Correlation (r)	Combination
NSE (Nash and Sutcliffe, 1970; Gupta et al., 2009)	$\frac{(\mu_S - \mu_O)}{\sigma_O}$	$\frac{(\sigma_S)}{(\sigma_O)}$	$r_{pearson}$	$2 * \alpha * r - \alpha^2 - \beta^2$
KGE (Gupta et al., 2009)	$\frac{(\mu_S)}{(\mu_O)}$			$1 - \sqrt{(\alpha - 1)^2 + (\beta - 1)^2 + (r - 1)^2}$
KGE* (modified version in Kling et al., 2012)		$\frac{[(\sigma_S)/(\mu_S)]}{[(\sigma_O)/(\mu_O)]}$		
NP (Pool et al., 2018)		$1 - \frac{1}{2}\sum_{i=1}^{n}\left\lvert\frac{x_{S,I(i)}}{n\mu_s} - \frac{x_{O,J(i)}}{n\mu_o}\right\rvert$	$r_{spearman}$	
SHE_g (gauged situation)	$\frac{[(\mu_S)/(\mu_P)]}{[(\mu_O)/(\mu_P)]}$	$\frac{[(\sigma_S)/(\sigma_P)]}{[(\sigma_O)/(\sigma_P)]}$		
SHE_u (ungauged situation with regionalized signatures)	$\frac{[(\mu_S)/(\mu_P)]}{[RR_{Pred}]}$	$\frac{[(\sigma_S)/(\sigma_P)]}{[VR_{Pred}]}$	$r^*_{spearman}$	

- S, O and P are simulated streamflow, and observed streamflow and precipitation, respectively.
- μ is the mean and σ is the standard deviation of streamflow.
- $x_{S,I(i)}$ is the simulated streamflow value where I(i) is the time step when the ith largest flow occurs within simulated time series and $x_{O,J(i)}$ is the observed streamflow value of target catchment where J(i) is the time step when the ith largest flow occurs within observed time series.
- VR_{Pred} and RR_{Pred} are regionalized variance ratio and runoff ratio for the target catchment derived using stepwise linear regression. Predictors of VR_{Pred} are aridity index, BFI-HOST and inland water percentage. Predictor of RR_{Pred} is only aridity index. Variance ratio is the ratio of standard deviation of streamflow to standard deviation of precipitation. Runoff ratio is the ratio of long-term mean of streamflow to long-term mean of precipitation.
- $r_{Pearson}$ = Pearson correlation between simulated and the observed streamflow in the target catchment
- $r_{Spearman}$ = Spearman rank correlation between simulated and the observed streamflow in the target catchment

$r'_{Spearman}$= Spearman rank correlation between simulated streamflow of a catchment which is assumed to be ungauged and the streamflow values obtained by inverse distance weighting interpolation of this catchment's three closest catchments' observed streamflow.

3.2 Application of SHE metric in ungauged catchments

Applying the SHE metric in ungauged situations requires estimates all of three metric components for ungauged basins. We perform this regionalization step in two different ways. Bias and variance components, i.e. runoff ratio and variance ratio, or related signatures have been widely regionalized using different types of regressions (e.g. Yadav et al., 2007; for GB).

We use the simplest and widely used strategy, stepwise linear regression, to establish the relationships between the catchment attributes and signatures (e.g. Almeida et al., 2016). We regionalize bias and variance signatures for 633 GB catchments testing 54 catchment attributes from CAMELS-GB representing topography, climate, hydrology, land cover, soils, hydrogeology and human influences (see Table S1 in supplemental information). Regionalized signatures are estimated using following procedure (see details in the supplementary material S1): (1) Stepwise regression is applied to each signature independently. Predictors are selected according to their p-values and the R^2 value of the resulting regression model. (2) 633 catchments are randomly divided into 5 groups. One group is left out each time and the remaining ones are used in the fitting of regression models for each signature (5-fold cross-validation). (3) After obtaining regression models in step 2 (see Table S2 and S3 in supplemental information), the signature values are estimated for the catchments in the group omitted during regression model development.

The correlation term is more complicated, given that we have no simple approach to regionalize a single value as is the case with the other two signatures. However, Archfield and Vogel (2010) have demonstrated that it is feasible to estimate correlation for ungauged locations using a geostatistical strategy. They introduced their map correlation method which selects the strongest correlated gauge as the reference gauge for an ungauged catchment, given that the nearest gauge was not always the most correlated one in their study of US catchments. The approach by Archfield and Vogel (2010) follows the basic idea of directly transferring streamflow from gauged to ungauged locations (see wider review of such approaches by He et al., 2011). Drogue and Plasse (2014) tested four different distance-based regionalization methods including the strategy by Archfield and Vogel (2010) for European catchments. They found that using multiple reference catchments rather than one is preferrable for assessing daily streamflow hydrographs in a densely gauged study domain. The simplest strategy to directly transfer streamflow is likely the one by Patil and Stieglitz (2012), who used inverse distance weighted (IDW) interpolation to transfer daily streamflow from multiple neighbouring gauged catchments to ungauged catchments in the US. Their approach is formulated as follows:

$$q(x) = \sum_{k=1}^{N} \frac{wk(x)}{\sum_{k=1}^{N} wk(x)} * q(xk) \text{ and } \mathrm{wk(x)} = \frac{1}{d(x,xk)^p}$$

where q(x) is daily streamflow (mm/day) at the ungauged catchment that is located at point x in the region, $q(x_k)$ is the daily streamflow of neighbouring reference catchment k located at point x_k in the region and N is the total number of neighbouring reference catchments for the interpolation. d is the distance between gauges of catchments and w is the interpolation weights of reference catchments. The exponent p is a positive real number, called a power parameter.

We adopt this approach for estimating streamflow to ungauged locations within our GB dataset because it works surprisingly well and because optimizing the regionalization performance is not our main concern. To identify a suitable number of reference catchments, we assume each catchment in turn to be ungauged, estimate the streamflow time series using IDW interpolation with different numbers of reference catchments (1-5 reference catchments), and calculate the Spearman Rank Correlation (SRC) between transferred and observed streamflow time series. We find that using three reference catchments provides optimum SRC estimate for the ungauged catchments in our sample (Figure S1 in supplemental information). We could actually use a similar streamflow transfer strategy to estimate the bias and variance terms but found this strategy to perform less well (see Figure S2 in supplemental information).

3.3 Rainfall-Runoff Model Implementation

We use a typical lumped parsimonious model structure widely used in Great Britain. The model structure, implemented in the Rainfall-Runoff Modelling Toolbox (RRMT; Wagener et al., 2001) combines a probability-distributed soil moisture accounting component (i.e. PDM), which represents the variability in soil moisture storage across a typical humid catchment using a distribution of storage depths (Moore, 2007), and a combination of two linear reservoirs in parallel for routing, one representing fast flow and the other representing slow flow (i.e. 2PAR), with a fixed split between them. Effective rainfall is produced as overflow from the PDM stores which are described as Pareto distribution based on two parameters, the maximum storage capacity, C_{max}, and parameter, b, describing the shape of the distribution. The effective rainfall (ER) is split with respect to parameter *a* describing the fraction of flow through the fast reservoir, while both reservoirs are defined by a single time constant (Wagener et al., 2001). The reason of choosing PDM is that it represents a flexibility in soil moisture accounting through its distribution function to influence the runoff response and combining it with 2PAR flow routing module provides different flow pathways for catchments across GB with different levels of baseflow contribution.

To calibrate the model, 10,000 parameter sets are independently sampled using uniform random sampling. The first 5% of the ten-year study period is used as a warm-up period. The parameter set producing the best performance according to SHE metric is used to obtain simulated streamflow. These numbers have been widely used in previous studies.

4 Results

First, we compare the values estimated for our SHE metric in gauged situations with previous efficiency metric implementations, i.e. KGE (Kling et al., 2012), NSE (Gupta et al., 2009) and NP (Pool et al., 2018). Figure 1 shows scatter plots where SHE values are correlated with KGE, NSE and NP values with Pearson correlation (i.e. PC) and Spearman rank correlation (i.e. SRC) values to varying degrees. Correlations are highest for SHE-NP (above 0.8), then SHE-KGE (around 0.8) and then SHE-NSE (0.6 to 0.67). Our formulation is most closely related to that of Pool et al. (2018) and Gupta et al (2009) due to the equal weighting of the terms within the efficiency metric.

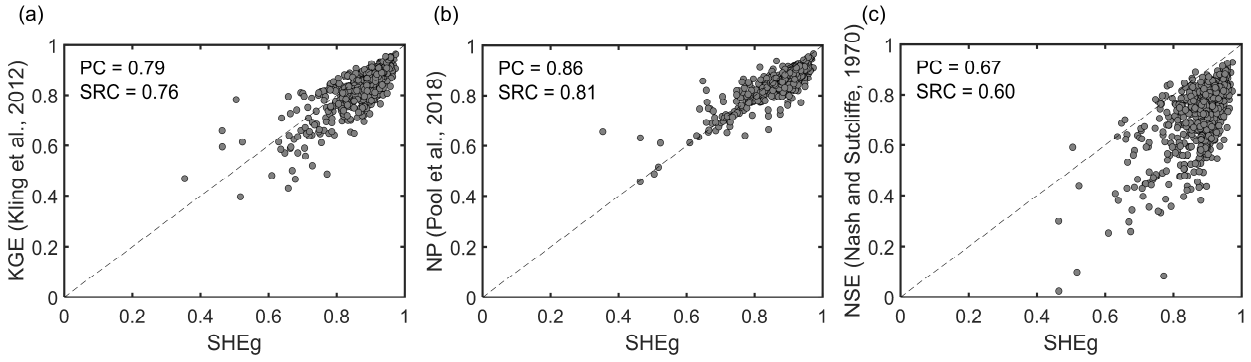

Figure 1. Scatter plots for (a) KGE vs. SHE, (b) NP vs. SHE and (c) NSE vs. SHE. x and y axes are limited to [0 1]. KGE, NP and NSE values are calculated using the best simulation values based on SHE metric values.

Second, we estimate the components of our metric for ungauged locations. The scatter plots in Figures 2a and 2b show that the predicted RR and VR using stepwise linear regression correlate well with observed RR and VR values. We find PC and SRC correlation values above 0.9. The maps indicate that predicted RR and VR values have similar patterns with decreases from the north-west to south-east of GB. As shown in Figure 2c for an estimate of correlation for ungauged locations, SRC values between observed and transferred streamflow values are above 0.8 for 94% of all catchments (77% above 0.9), even when using the simple inverse distance method with the three closest catchments. All components of our SHE metric can therefore be estimated individually in ungauged catchments within our study domain.

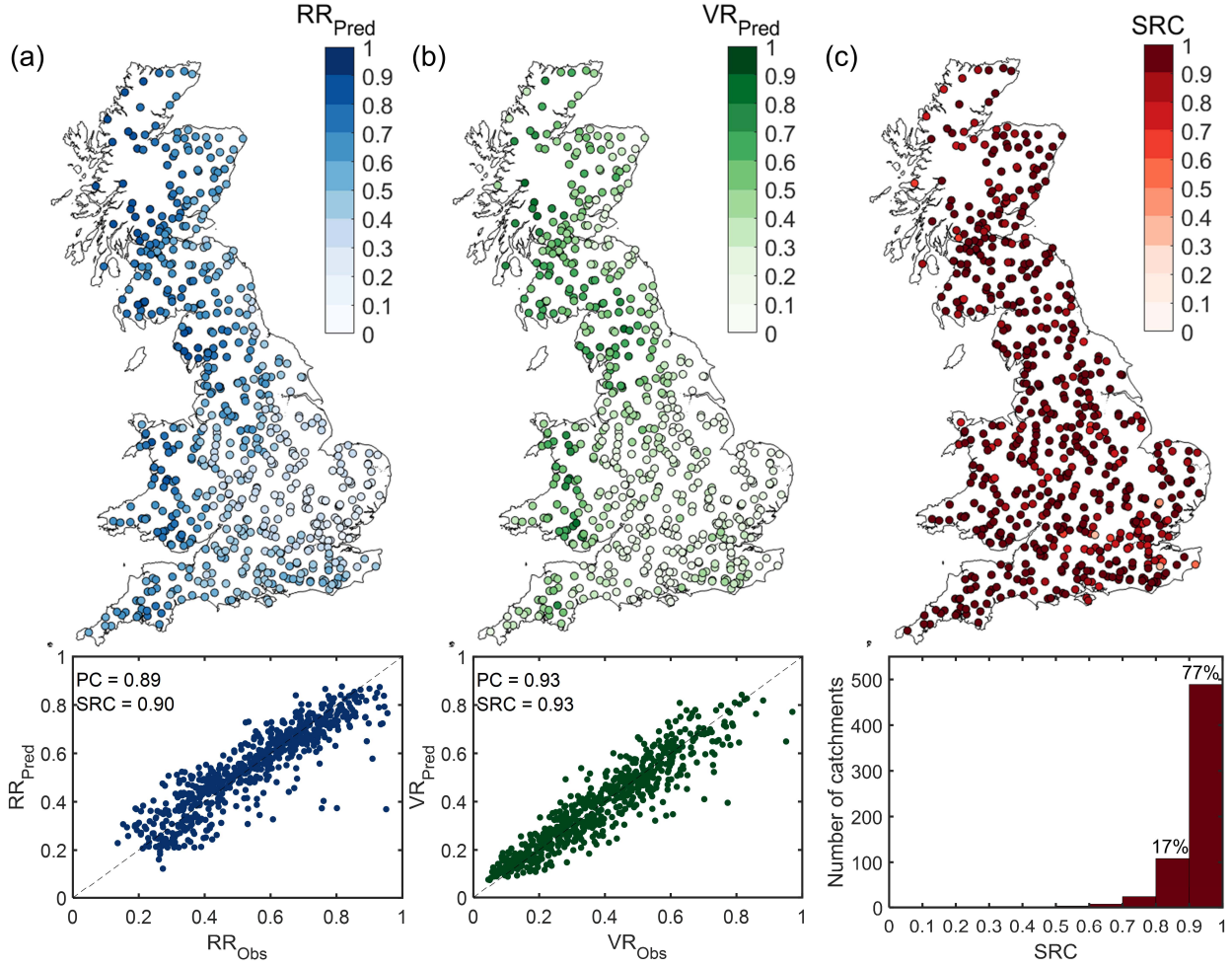

Figure 2. (a) Predicted RR map and scatter plot for predicted vs. observed RR, (b) predicted VR map and scatter plot for predicted vs. observed VR and (c) map illustrating SRC values between observed streamflow of catchments and the streamflow values calculated by taking inverse distance interpolation of their closest three catchments' observed streamflows and its histogram plot. Predictor of RR is aridity index and predictors of VR are aridity index, BFI-HOST and inland water percentage.

And third, we calculate the differences between SHE values for gauged and ungauged cases to evaluate how well we can estimate the performance of a model for ungauged catchments, in contrast to gauged catchments. Figures 3a, 3b and 3c shows histograms of the differences between SHE values for gauged and ungauged cases (i.e. $SHE_g - SHE_u$). Cumulative

distribution functions (CDF) plots of the individual difference values are color-coded by (a) bias component difference (i.e. Δβ), (b) variance component difference (i.e. Δα) and (c) correlation component difference (i.e. Δr). The histograms (all three are identical) show that more than 50% of 633 catchments have difference values between -0.1 and 0.1, while 78% of them have difference values between -0.2 and 0.2. Low values of SHE difference are associated with small differences in the bias, variance, and correlation terms (see CDF plots in Figure 3). CDF plots also show that catchments with high positive differences (i.e. >0.3) have the highest positive and the lowest negative values of the bias and variance component differences, respectively, suggesting the poor regionalization is a problem there. Figure 3c shows that correlation component differences are overall very small across catchments except for very few catchments with high positive differences. In summary, the results imply that when the regionalization of the bias and variance signatures works, we can obtain similar SHE values for both gauged and ungauged cases.

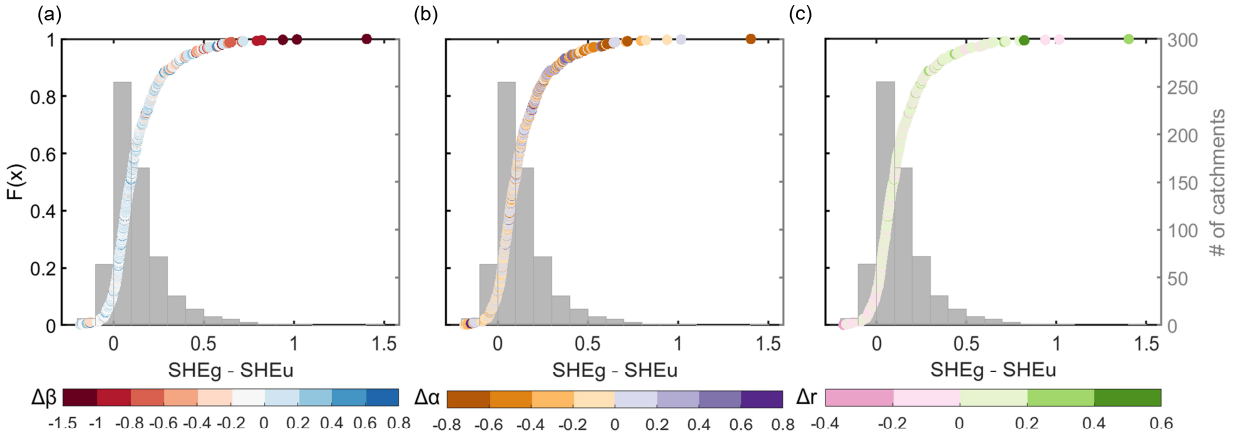

Figure 3. Cumulative distribution function (i.e. cdf) plot and histogram plot of difference between SHE for gauged and ungauged cases (i.e. SHE_g – SHE_u). Cdf plot is color-coded by (a) bias component difference (Δβ), (b) variance component difference (Δα) and (c) correlation component difference (Δr) between SHE formulations for gauged and ungauged cases summarized in Table 1.

5 Discussion and Conclusions

In summary, we introduced a new signature-based hydrologic efficiency (SHE) metric based on the idea that a model's fit to signatures will be easier to interpret hydrologically, and more importantly, that we can estimate it directly in ungauged basins. The SHE metric is correlated to different degree with existing metrics, and we show how its components, and hence the metric itself, can be estimated in ungauged catchments.

A flexible efficiency metric based on signatures provides significant opportunity for hydrologically relevant diagnostic model calibration and evaluation (Yadav et al., 2007; Yilmaz et al., 2008; Shafii and Tolson, 2015). Here, we simply replace the statistical components of the KGE (Gupta et al., 2009) with signatures suitable for our study domain, Great Britain. We chose to use runoff ratio and variance ratio as our signatures to represent bias and variance aspects of the hydrograph. However, other signatures could and should be considered for different study

domains. Hydrologists have investigated many signatures and found different ones to be useful to characterize major hydrologic functions or hydrograph aspects of catchments depending on the study domain (McMillan, 2020). Different aspects of the flow duration curve have for example been used to characterize the variability of flow through different signatures (e.g. Yilmaz et al., 2008; Sawicz et al., 2011; Westerberg et al., 2011; Pool et al., 2018; McMillan, 2021). It might be useful to use different signatures depending on whether study domains for example contain catchments with significant snow or those in arid domains.

We do not believe that SHE would be universally applicable in this form everywhere in the world. Actually, we believe that the different components should be replaced by appropriate signatures of a catchment's, water balance, its damping, and its translation of precipitation variability into streamflow variability and timing. Different signatures might be best suited to represent these components depending on whether the study domain is for example located in a temperate, dry or cold part of the world. Equally, existing regionalized streamflow indices correlated with these components might provide a baseline from which such a metric can be estimated in both gauged and ungauged catchments. An advantage of this opportunity and need for tailoring is that making these choices puts the discussion about suitable objective functions into the realm of hydrology, rather than just statistics.

The issue of signature choice is also linked to the ability for regionalising signatures or indices correlated with the components of the efficiency metric. Many regionalisation studies exist (e.g. He et al., 2011; Wagener and Montanari, 2011), though in how far these studies provide a regional basis to calculate efficiency metrics from in ungauged locations has so far been unexplored. One issue we did not tackle here in this context is that of uncertainty in these regionalisation estimates (e.g. Zhang et al., 2008; Kapangaziwiri et al., 2012; Westerberg et al., 2014). Uncertainties originate from the underlying measurements of physical catchment properties and of hydro-meteorological variables, from processing of the original observations, and from choices made regarding space-time averaging etc. (McMillan et al., 2022; Westerberg et al., 2016). There is opportunity for integrating uncertainty in a coherent statistical framework covering both gauged and ungauged situations, which should significantly increase the value of available regionalised information in the context of model calibration and evaluation.

Acknowledgments

MK was funded by Ministry of National Education, the Republic of Turkey. Partial support for GC was provided by a NERC grant NE/V009060/1 and UKRI Future Leaders Fellowship award [MR/V022857/1]. Funding for TW has been provided by the Alexander von Humboldt Foundation in the framework of the Alexander von Humboldt Professorship endowed by the German Federal Ministry of Education and Research.

Open Research

Full period of record daily rainfall, streamflow, potential evapotranspiration for all our daily flow gauging stations throughout the GB and and catchment attributes are taken from CAMELS-GB dataset (Coxon et al., 2020) and available at https://doi.org/10.5285/8344e4f3-d2ea-44f5-8afa-

86d2987543a9. BFI-HOST of each catchment is obtained from NRFA website (https://nrfa.ceh.ac.uk/data/search) where detailed information of each stream gauges is given. SHE (for both gauged and ungauged conditions), KGE, NP and NSE values obtained for 659 GB catchments and their simulated streamflow values for the best simulations obtained by this study will also be made available in the Bristol data repository.

References

Almeida, S., Le Vine, N., McIntyre, N., Wagener, T., & Buytaert, W. (2016). Accounting for dependencies in regionalized signatures for predictions in ungauged catchments. *Hydrology and Earth System Sciences*, *20*(2), 887-901.

Archfield, S. A., & Vogel, R. M. (2010). Map correlation method: Selection of a reference streamgage to estimate daily streamflow at ungaged catchments. *Water resources research*, *46*(10).

Bulygina, N., McIntyre, N., & Wheater, H. (2009). Conditioning rainfall-runoff model parameters for ungauged catchments and land management impacts analysis. *Hydrology and Earth System Sciences*, *13*(6), 893-904.

Clark, M.P., Vogel, R.M., Lamontagne, J.R., Mizukami, N., Knoben, W.J.M., Tang, G. et al. (2021). The abuse of popular performance metrics in hydrologic modelling. *Water Resources Research*, *57*(9), e2020WR029001

Coxon, G., Addor, N., Bloomfield, J. P., Freer, J., Fry, M., Hannaford, J., Howden, N. J. K., Lane, R., Lewis, M., Robinson, E. L., Wagener, T., & Woods, R. (2020). CAMELS-GB: hydrometeorological time series and landscape attributes for 671 catchments in Great Britain. *Earth System Science Data*, *12*(4), 2459-2483.

354 Drogue, G. P., & Plasse, J. (2014). How can a few streamflow measurements help to predict
355 daily hydrographs at almost ungauged sites? *Hydrological Sciences Journal*, *59*(12), 2126-2142.
356 Gershenfeld, N. (1999). The nature of mathematical modelling. Cambridge University Press.
357 Giani, G., Rico-Ramirez, M. A., & Woods, R. A. (2021). A practical, objective, and robust
358 technique to directly estimate catchment response time. *Water Resources Research*, *57*(2),
359 e2020WR028201.
360 Gupta, H. V., Wagener, T., & Liu, Y. (2008). Reconciling theory with observations: elements of
361 a diagnostic approach to model evaluation. *Hydrological Processes: An International*
362 *Journal*, *22*(18), 3802-3813.
363 Gupta, H. V., Kling, H., Yilmaz, K. K., & Martinez, G. F. (2009). Decomposition of the mean
364 squared error and NSE performance criteria: Implications for improving hydrological
365 modelling. *Journal of hydrology*, *377*(1-2), 80-91.
366 Guo, Y., Zhang, Y., Zhang, L., & Wang, Z. (2021). Regionalization of hydrological modeling for
367 predicting streamflow in ungauged catchments: A comprehensive review. *Wiley Interdisciplinary*
368 *Reviews: Water*, *8*(1), e1487.
369 He, Y., Bárdossy, A., & Zehe, E. (2011). A review of regionalisation for continuous streamflow
370 simulation. *Hydrology and Earth System Sciences*, *15*(11), 3539-3553.
371 Hrachowitz, M., Fovet, O., Ruiz, L., Euser, T., Gharari, S., Nijzink, R., ... & Gascuel-Odoux, C.
372 (2014). Process consistency in models: The importance of system signatures, expert knowledge,
373 and process complexity. Water resources research, 50(9), 7445-7469.
374 Kapangaziwiri, E., Hughes, D. A., & Wagener, T. (2012). Incorporating uncertainty in
375 hydrological predictions for gauged and ungauged basins in southern Africa. *Hydrological*
376 *Sciences Journal*, *57*(5), 1000-1019.

377 Kling, H., Fuchs, M., & Paulin, M. (2012). Runoff conditions in the upper Danube basin under
378 an ensemble of climate change scenarios. *Journal of hydrology*, *424*, 264-277.
379 Knoben, W. J., Freer, J. E., & Woods, R. A. (2019). Inherent benchmark or not? Comparing
380 Nash–Sutcliffe and Kling–Gupta efficiency scores. *Hydrology and Earth System*
381 *Sciences*, *23*(10), 4323-4331.
382 McMillan, H. (2020). Linking hydrologic signatures to hydrologic processes: A
383 review. *Hydrological Processes*, *34*(6), 1393-1409.
384 McMillan, H. K. (2021). A review of hydrologic signatures and their applications. *Wiley*
385 *Interdisciplinary Reviews: Water*, *8*(1), e1499.
386 McMillan, H. K., Coxon, G., Sikorska-Senoner, A. E., & Westerberg, I. K. (2022). Impacts of
387 observational uncertainty on analysis and modelling of hydrological processes: Preface.
388 Hydrological Processes, 36 (2),[e14481]. https://doi. org/10.1002/hyp. 14481.
389 Milly, P. C. D. (1994). Climate, soil water storage, and the average annual water balance. *Water*
390 *Resources Research*, *30*(7), 2143-2156.
391 Moges, E., Ruddell, B. L., Zhang, L., Driscoll, J. M., Norton, P., Perez, F., & Larsen, L. G.
392 (2022). HydroBench: Jupyter supported reproducible hydrological model benchmarking and
393 diagnostic tool. *Frontiers in Earth Science*, 1469.
394 Moore, R. J. (2007). The PDM rainfall-runoff model. *Hydrology and Earth System*
395 *Sciences*, *11*(1), 483-499.
396 Moriasi, D. N., Arnold, J. G., Van Liew, M. W., Bingner, R. L., Harmel, R. D., & Veith, T. L.
397 (2007). Model evaluation guidelines for systematic quantification of accuracy in watershed
398 simulations. *Transactions of the ASABE*, *50*(3), 885-900.

399 Murphy, A. H. (1988). Skill scores based on the mean square error and their relationships to the
400 correlation coefficient. *Monthly weather review*, *116*(12), 2417-2424.
401 Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I—
402 A discussion of principles. *Journal of hydrology*, *10*(3), 282-290.
403 Olden, J. D., & Poff, N. L. (2003). Redundancy and the choice of hydrologic indices for
404 characterizing streamflow regimes. *River research and applications*, *19*(2), 101-121.
405 Patil, S., & Stieglitz, M. (2012). Controls on hydrologic similarity: role of nearby gauged
406 catchments for prediction at an ungauged catchment. *Hydrology and Earth System*
407 *Sciences*, *16*(2), 551-562.
408 Pool, S., Vis, M., & Seibert, J. (2018). Evaluating model performance: towards a non-parametric
409 variant of the Kling-Gupta efficiency. *Hydrological Sciences Journal*, *63*(13-14), 1941-1953.
410 Pool, S., Vis, M., & Seibert, J. (2021). Regionalization for ungauged catchments—lessons
411 learned from a comparative large-sample study. *Water Resources Research*, *57*(10),
412 e2021WR030437.
413 Rogelis, M. C., Werner, M., Obregón, N., & Wright, N. (2016). Hydrological model assessment
414 for flood early warning in a tropical high mountain basin. *Hydrology and Earth System Sciences*
415 *Discussions*, 1-36.
416 Sankarasubramanian, A., Vogel, R. M., & Limbrunner, J. F. (2001). Climate elasticity of
417 streamflow in the United States. *Water Resources Research*, *37*(6), 1771-1781.
418 Sawicz, K., Wagener, T., Sivapalan, M., Troch, P. A., & Carrillo, G. (2011). Catchment
419 classification: empirical analysis of hydrologic similarity based on catchment function in the
420 eastern USA. *Hydrology and Earth System Sciences*, *15*(9), 2895-2911.

Schaefli, B., & Gupta, H. V. (2007). Do Nash values have value?. *Hydrological processes*, *21*(ARTICLE), 2075-2080.

Schwemmle, R., Demand, D., & Weiler, M. (2021). Diagnostic efficiency–specific evaluation of model performance. *Hydrology and Earth System Sciences*, *25*(4), 2187-2198.

Seibert, J. (2001). On the need for benchmarks in hydrological modelling. *Hydrological Processes*, *15*(6), 1063-1064.

Seibert, J., Vis, M. J., Lewis, E., & Meerveld, H. V. (2018). Upper and lower benchmarks in hydrological modelling. *Hydrological processes*, 32(8), 1120-1125.

Shafii, M., & Tolson, B. A. (2015). Optimizing hydrological consistency by incorporating hydrological signatures into model calibration objectives. *Water Resources Research*, *51*(5), 3796-3814.

Towner, J., Cloke, H. L., Zsoter, E., Flamig, Z., Hoch, J. M., Bazo, J., Coughlan de Perez, E., & Stephens, E. M. (2019). Assessing the performance of global hydrological models for capturing peak river flows in the Amazon basin. *Hydrology and Earth System Sciences*, *23*(7), 3057-3080.

Vogel, R. M., Wilson, I., & Daly, C. (1999). Regional regression models of annual streamflow for the United States. *Journal of Irrigation and Drainage Engineering*, *125*(3), 148-157.

Wagener, T., Lees, M. J., & Wheater, H. S. (2001). A toolkit for the development and application of parsimonious hydrological models. *Mathematical models of large watershed hydrology*, *1*, 87-136.

Wagener, T., & Montanari, A. (2011). Convergence of approaches toward reducing uncertainty in predictions in ungauged basins. *Water Resources Research*, *47*(6).

Węglarczyk, S. (1998). The interdependence and applicability of some statistical quality measures for hydrological models. *Journal of Hydrology*, *206*(1-2), 98-103.

444 Westerberg, I. K., Guerrero, J. L., Younger, P. M., Beven, K. J., Seibert, J., Halldin, S., Freer, J.
445 R., & Xu, C. Y. (2011). Calibration of hydrological models using flow- duration curves.
446 *Hydrology and Earth System Sciences*, 15(7), 2205-2227.
447 Westerberg, I. K., Gong, L., Beven, K. J., Seibert, J., Semedo, A., Xu, C. Y., & Halldin, S.
448 (2014). Regional water balance modelling using flow-duration curves with observational
449 uncertainties. *Hydrology and Earth System Sciences*, *18*(8), 2993-3013.
450 Westerberg, I. K., Wagener, T., Coxon, G., McMillan, H. K., Castellarin, A., Montanari, A., &
451 Freer, J. (2016). Uncertainty in hydrological signatures for gauged and ungauged
452 catchments. *Water Resources Research*, *52*(3), 1847-1865.
453 Yadav, M., Wagener, T., & Gupta, H. (2007). Regionalization of constraints on expected
454 watershed response behavior for improved predictions in ungauged basins. *Advances in water*
455 *resources*, *30*(8), 1756-1774.
456 Yilmaz, K. K., Gupta, H. V., & Wagener, T. (2008). A process-based diagnostic approach to
457 model evaluation: Application to the NWS distributed hydrologic model. *Water Resources*
458 *Research*, *44*(9).
459 Zhang, Z., Wagener, T., Reed, P., & Bhushan, R. (2008). Reducing uncertainty in predictions in
460 ungauged basins by combining hydrologic indices regionalization and multiobjective
461 optimization. Water Resources Research, 44(12).
462