Revista Brasileira de Recursos Hídricos
Brazilian Journal of Water Resources
Versão On-line ISSN 2318-0331
RBRH, Porto Alegre, v. 25, e42, 2020
Technical Note

https://doi.org/10.1590/2318-0331.252020190155

Objective functions used as performance metrics for hydrological models: state-of-the-art and critical analysis

Funções-objetivo utilizadas como medidas de desempenho de modelos hidrológicos: estado-da-arte e análise crítica

Paloma Mara de Lima Ferreira[1], **Adriano Rolim da Paz**[1] **& Juan Martín Bravo**[2]

[1]Universidade Federal da Paraíba, João Pessoa, PB, Brasil
[2]Instituto de Pesquisas Hidráulicas, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brasil
E-mails: paloma_mara@hotmail.com.br (PMLF), adrianorpaz@yahoo.com.br (ARP), jumarbra@iph.ufrgs.br (JMB)

Received: October 23, 2019 - Revised July 07, 2020 - Accepted July 19, 2020

ABSTRACT

Hydrological models (HMs) can be applied for different purposes, and a key step is model calibration using objective functions (OF) to quantify the agreement between observed and calculated discharges. Fully understanding the OF is important to properly take advantage of model calibration and interpret the results. This study evaluates 36 OF proposed in the literature, considering two watersheds of different hydrological regimes. Daily simulated streamflow time-series, using a distributed hydrological model (MGB-IPH), and ten daily streamflow synthetic time-series, generated from the observed and calculated streamflows, were used in the analysis of each watershed. These synthetic data were used to evaluate how does each metric evaluate hypothetical cases that present isolated very well known error behaviors. Despite of all NSE-derived (Nash-Sutcliffe efficiency) metrics that use the square of the residuals in their formulation have shown higher sensitivity to errors in high flows, the ones that use daily and monthly averages of flow rates in absolute terms were more stringent than the others to assess HMs performance. Low flow errors were better evaluated by metrics that use the flow logarithm. The constant presence of zero flow rates deteriorate them significantly, with the exception of the metrics TRMSE (Transformed root mean square error) did not demonstrate this problem. An observed limitation of the formulations of some metrics was that the errors of overestimation or underestimation are compensated. Our results reassert that each metric should be interpreted specifically thinking about the aspects it has been proposed for, and simultaneously taking into account a set of metrics would lead to a broader evaluation of HM ability (e.g. multiobjective model evaluation). We recommend that the use of synthetic time series as those proposed in this work could be useful as an auxiliary step towards better understanding the evaluation of a calibrated hydrological model for each study case, taking into account model capabilities and observed hydrologic regime characteristics.

Keywords: Model calibration; Hydrologic simulation; Performance measures; MGB-IPH.

RESUMO

Modelos hidrológicos (MHs) podem ser aplicados para diferentes propósitos, uma etapa fundamental é a calibração do modelo usando funções objetivo (FO) para quantificar a concordância entre as vazões observadas e calculadas. O entendimento completo das FO é importante para aproveitar adequadamente a calibração do modelo e interpretar os resultados. Este estudo avalia 36 FO propostas na literatura, considerando duas bacias hidrográficas de diferentes regimes hidrológicos. Séries temporais diárias de vazão simulada, usando um modelo hidrológico distribuído (MGB-IPH), e dez séries temporais sintéticas diárias, geradas a partir das vazões observadas e calculadas, foram usadas na análise de cada bacia hidrográfica. Esses dados sintéticos foram usados para avaliar como cada métrica avalia os casos hipotéticos que apresentam comportamentos de erro isolados muito conhecidos. Apesar de todas as métricas derivadas de NSE (eficiência de Nash-Sutcliffe) que usam o quadrado dos resíduos em sua formulação terem demonstrado maior sensibilidade a erros nas vazões altas, os que usam médias diárias e mensais das vazões em termos absolutos foram mais rigorosos que os outros para avaliar o desempenho dos MHs. Erros nas vazões baixas foram melhor avaliados por métricas que usam o logaritmo das vazões. A presença constante de vazões zero os deteriora significativamente, com exceção das métricas TRMSE (erro quadrático médio da raiz transformada) não demonstraram esse problema. Uma limitação observada das formulações de algumas métricas foi que os erros de superestimação ou subestimação são compensados. Nossos resultados reafirmam que cada métrica deve ser interpretada pensando especificamente nos aspectos para os quais foi proposta e, considerando simultaneamente um conjunto de métricas, levaria a uma avaliação mais ampla da capacidade do MH (ex: avaliação de modelo multiobjetivo). Recomendamos que o uso de séries temporais sintéticas, como as propostas neste trabalho, possa ser útil como um passo auxiliar no melhor entendimento da avaliação de um modelo hidrológico calibrado para cada estudo de caso, levando em consideração as capacidades do modelo e as características observadas do regime hidrológico.

Palavras-chave: Calibração de modelo; Simulação hidrológica; Medidas de desempenho; MGB-IPH.

INTRODUCTION

Hydrologic models (HMs) are used to represent hydrological processes in order to obtain information for water resources planning and management. They enable a rapid response to several scenarios and assist decision-making processes regarding land use change, climate variability, and water-intensive scenarios, among others, for water resources in a given region (Tucci, 2005; Beven, 2012).

HMs usually need to be calibrated to be useful in solving practical problems. During the calibration process, parameter values are defined to enable the model to closely match the behavior of the real-world system. Model calibration partially compensates for different types of hydrological uncertainties, such as those associated with input data, hydrological processes, mathematical formulation of the hydrologic model, spatiotemporal discretization, and observations (Efstratiadis & Koutsoyiannis, 2010; Beven, 2012).

The response of the hydrological system is commonly represented by observed streamflow time-series. Thus, during calibration of HMs, observed and calculated hydrographs are compared at points along the drainage network. More recently, efforts have been made to combine such analysis together with other hydrological process variables such as evapotranspiration (Herman et al., 2018), soil moisture (Rajib et al., 2016) and surface temperature (Zink et al., 2018). However, comparison between observed (*Qo*) and calculated (*Qc*) streamflows predominates as the most widely used approach (Troin et al., 2015; Zhang et al., 2016; Molina-Navarro et al., 2017).

One approach for HM calibration is manual calibration, based on manually changing model parameters and visually comparing observed and calculated hydrographs. This is an intuitive way to judge the fit quality and is even preferred by many users (Pappenberger & Beven, 2004), being actually the most widely used one (Boyle et al., 2000). This procedure uses the experience of the hydrologist to assess several aspects of observed and calculated hydrograph similarities, such as peak flows, peak times, rise and recession limbs, drought flows, and flood durations. However, subjectivity in choosing one of many different parameter sets results from personal preferences for denoting more the peak flow or drought errors (Krause et al., 2005; Garcia et al., 2017), even when a model that represents the overall behavior of the observed hydrographs is intended. Another remarkable shortcoming is that the manual search for optimal parameters poorly explores the parameter space.

Automatic calibration is a second approach for HM calibration. It uses metrics to mathematically assess the degree of agreement between *Qo* and *Qc*. Each metric weights the error between *Qo* and *Qc*, considering a specific mathematical formulation that must be minimized or maximized as an objective function (OF) of an optimization problem. Manual calibration could also be performed by manually varying model parameters and evaluating model performance by inspecting such metrics.

Metrics such as correlation coefficient (*r*), coefficient of determination (r^2), root mean square error (RMSE), and Nash-Suttclife efficiency (NSE) are the most widely used (Gupta et al., 2009; Westerberg et al., 2011; Wohling et al., 2013). Coefficients such as *r* and r^2 evaluate the collinearity between *Qo* and *Qc*, while metrics such as *RMSE* measure the mean error between *Qc* and *Qo* in the flow unit itself. Metrics such as NSE assess the HM performance against a baseline model represented by the mean of all streamflow observations.

As each metric weights the error between *Qo* and *Qc* in different ways, its formulation and selection criteria should be considered for the correct interpretation of results. An HM may be applied for different purposes, which means that the ability of an HM to reproduce different aspects of the observed streamflow regime may vary in relevance for a given application (Garcia et al., 2017). For example, an HM developed for estimating water availability in semiarid climate regions should be evaluated for its ability to represent the drought period. On the other hand, an HM for flood warning must be evaluated regarding its capability to simulate high streamflows.

The use of HMs is increasing mainly due to the development of user-friendly interfaces, the integration and automation of data preparation steps within Geographic Information Systems, and the inclusion of automatic calibration modules. All of this speeds up the application of HMs, but it means that less attention and time is dedicated to critical appraisal of the data, evaluation of the calibration process, and analysis of overall HM results. One of the usually neglected steps is ensuring the correct selection of OFs for HM calibration. The calibration process requires other issues to be addressed, such as the size of observed streamflow time-series (e.g. Li et al., 2010; Nelson et al., 2017), the mathematical method of searching for the optimal parameters set (e.g. Bravo et al., 2009), and the computational cost involved (Gutierrez et al., 2019).

In the literature, dozens of metrics are used as OFs for HM calibration (e.g. Legates & McCabe Junior, 1999; Krause et al., 2005; Moriasi et al., 2007; Gupta et al., 2009; Muleta, 2012; Romanowicz et al., 2013; Wohling et al., 2013; Fowler et al., 2018). This large number of alternatives contrasts with the repeated use of a small set of metrics in current HM calibration approaches. Often, such use is made without criteria, which may lead to mistaken conclusions about the HM performance.

This study assesses 36 metrics that have been proposed in literature for HM calibration by comparing calculated and observed hydrographs.

Each metric selected for this study has its specific formulation that differs from the others and has been explicitly adopted in one or more model calibration applications according to the mentioned references. But indeed some metrics present strong similarities among them. This review of metrics is exactly one of the contributions of this research. Moreover, the similarities or differences obtained within our results may help readers to better understand which metrics work similar to each other.

In addition, an analysis of each metric is carried out in order to verify how it is influenced by errors in several components of the calculated hydrographs (e.g. errors in the drought season, errors in the rainy season, magnitude of the error). Ten synthetic streamflow time-series were generated to be tested and evaluated by each metric, in order to see how does each one evaluate hypothetical cases that present isolated very well known error behaviors. Two Brazilian large-scale watersheds with contrasting characteristics (perennial vs intermittent streamflows) form the case study.

METRICS AS OBJECTIVE FUNCTIONS IN HYDROLOGIC MODEL CALIBRATION

A total of 36 metrics commonly used as OFs for HM calibration were identified and selected from an extensive literature review (Table 1). This list cannot be considered exhaustive, and other metrics not included in the list were used in specific analyses during HM calibration (e.g. the Richard-Bark flashness index proposed by Parker et al., 2019).

Each OF listed in Table 1 is presented with its mathematical formulation and its minimum, maximum and optimal values. The following section discusses the main issues related to each OF, presenting several references for further details.

Metrics *r* and *r*^{2} are some of the most commonly used in several scientific areas and evaluate the degree of linear association and dispersion between two datasets (e.g. *Qo* and *Qc*).

NSE is one of the most widespread OFs adopted for HM calibration. Metrics such as NSE assess the HM performance

Table 1. Metrics used to assess the performance of hydrological models.

Name (Symbology)	Mathematical formulation	Min, Max, Optimal	Units	References
Linear correlation coefficient (r)	$\frac{\sum_{i=1}^{n}\left(Q_{o(i)}-\overline{Q_{o(i)}}\right)\left(Q_{c(i)}-\overline{Q_{c(i)}}\right)}{\sqrt{\sum_{i=1}^{n}\left(Q_{o(i)}-\overline{Q_{o(i)}}\right)^2 * \sum_{i=1}^{n}\left(Q_{c(i)}-\overline{Q_{c(i)}}\right)^2}}$	-1, 1, 1	-	Wohling et al. (2013)
Coefficient of determination (r^2)	$\left(\frac{\sum_{i=1}^{n}\left(Q_{o(i)}-\overline{Q_{o(i)}}\right)\left(Q_{c(i)}-\overline{Q_{c(i)}}\right)}{\sqrt{\sum_{i=1}^{n}\left(Q_{o(i)}-\overline{Q_{o(i)}}\right)^2 * \sum_{i=1}^{n}\left(Q_{c(i)}-\overline{Q_{c(i)}}\right)^2}}\right)^2$	0, 1, 1	-	Romanowicz et al.(2013)
Nash-Sutcliffe efficiency (NSE)	$1-\frac{\sum_{i=1}^{n}\left(Q_{c(i)}-Q_{o(i)}\right)^2}{\sum_{i=1}^{n}\left(Q_{o(i)}-\overline{Q_{o(i)}}\right)^2}$	-∞, 1, 1	-	Wohling et al. (2013)
NSE on log transformed daily flows (LNS)	$1-\frac{\sum_{i=1}^{n}\left(logQ_{c(i)}-logQ_{o(i)}\right)^2}{\sum_{i=1}^{n}\left(logQ_{o(i)}-\overline{logQo}_{(i)}\right)^2}$	-∞, 1, 1	-	Romanowicz et al. (2013)
Modified forms of NSE (MNS)	$1-\frac{\sum_{i=1}^{n}\left\lvert Q_{c(i)}-Q_{o(i)}\right\rvert}{\sum_{i=1}^{n}\left\lvert Q_{o(i)}-\overline{Q_{o(i)}}\right\rvert}$	-∞, 1, 1	-	Muleta (2012)
NSE with calendar day mean (NSD)	$1-\frac{\sum_{i=1}^{n}\left(Q_{c(i)}-Q_{o(i)}\right)^2}{\sum_{i=1}^{n}\left(Q_{o(i)}-\bar{Q}_D\right)^2}$	-∞, 1, 1	-	Muleta (2012)
NSE with calendar day mean calculated on log transformed daily flows (LNSD)	$1-\frac{\sum_{i=1}^{n}\left(ln\,Q_{c(i)}-ln\,Q_{o(i)}\right)^2}{\sum_{i=1}^{n}\left(ln\,Q_{o(i)}-ln\,\bar{Q}_D\right)^2}$	-∞, 1, 1	-	Muleta (2012)
Modified form of NSE with calendar day mean (MNSD)	$1-\frac{\sum_{i=1}^{n}\left\lvert Q_{c(i)}-Q_{o(i)}\right\rvert}{\sum_{i=1}^{n}\left\lvert Q_{o(i)}-\bar{Q}_D\right\rvert}$	-∞, 1, 1	-	Muleta (2012)

Legend: Qo(i) and Qc(i) are the observed and calculated daily streamflow at time-interval i, $\bar{Q}_{o(i)}$ and $\bar{Q}_{c(i)}$ are the observed and calculated mean streamflows, mQo(i) and mQc(i) are the total monthly streamflows observed and calculated, $\overline{Q_D}$ is the interannual calendar day mean observed streamflows, $Q_{ref(i)}$ is the average reference streamflows, $\hat{Q}_{o(i)}$ and $\hat{Q}_{c(i)}$ are the observed and calculated transformed streamflows, *K* is the total number of years in the time-series, ω is a weighting parameter (used ω = 0.1), $QD_c(p)$ and $QD_o(p)$ are the observed and calculated streamflow for the probability p of the duration curve, Δp is the interval used from the duration curve for the sum, and n is the total number of records in the time-series.

Table 1. Continued...

Name (Symbology)	Mathematical formulation	Min, Max, Optimal	Units	References
NSE with calendar monthly mean as reference model (NSM)	$1-\frac{\sum_{i=1}^{n}\left(Q_{c(i)}-Q_{o(i)}\right)^{2}}{\sum_{i=1}^{n}\left(Q_{o(i)}-\bar{Q}_{ref}\right)^{2}}$	-∞, 1, 1	-	Schaefli And Gupta (2007)
Persistence Index (PI)	$1-\frac{\sum_{i=1}^{n}\left(Q_{c(i)}-Q_{o(i)}\right)^{2}}{\sum_{i=1}^{n}\left(Q_{o(i)}-Q_{o(i-1)}\right)^{2}}$	-∞, 1, 1	-	Gupta et al. (1999)
High flow (HF)	$1-\frac{\sum_{i=1}^{n}\left(Q_{o(i)}+\overline{Q_{o(i)}}\right)\left(Q_{c(i)}-Q_{o(i)}\right)^{2}}{\sum_{i=1}^{n}\left(Q_{o(i)}+\overline{Q_{o(i)}}\right)\left(Q_{o(i)}-\overline{Q_{o(i)}}\right)^{2}}$	-∞, 1, 1	-	Rwetabula et al. (2012)
Index of agreement (D)	$1-\left(\frac{\sum_{i=1}^{n}\left(Q_{o(i)}-Q_{c(i)}\right)^{2}}{\sum_{i=1}^{n}\left(\left\|Q_{c(i)}-\overline{Q_{o(i)}}\right\|+\left\|Q_{o(i)}-\overline{Q_{o(i)}}\right\|\right)^{2}}\right)$	0, 1, 1	-	Muleta (2012)
Relative variability (α)	$\sum_{i=1}^{n}\left(Q_{c(i)}-\overline{Q_{c(i)}}\right)/\sum_{i=1}^{n}\left(Q_{o(i)}-\overline{Q_{o(i)}}\right)$	0, ∞, 1	-	Wohling et al. (2013)
Normalised bias of flows (β)	$\overline{Q_{c(i)}}-\overline{Q_{o(i)}}/\sqrt{\frac{1}{n}*\sum_{i=1}^{n}\left(Q_{o(i)}-\overline{Q_{o(i)}}\right)^{2}}$	-∞, 1, 0	-	Wohling et al. (2013)
Kling-Gupta efficiency (KGE)	$1-\sqrt{(r-1)^{2}+(\alpha-1)^{2}+\left(\frac{\overline{Q_{c(i)}}}{\overline{Q_{o(i)}}}-1\right)^{2}}$	-∞, 1, 1	-	Wohling et al. (2013)
Mean error (ME)	$\frac{1}{n}\sum_{i=1}^{n}\left(Q_{c(i)}-Q_{o(i)}\right)$	-∞, ∞, 0	m^3/s	Wohling et al. (2013)
Mean absolute error (MAE)	$\frac{1}{n}\sum_{i=1}^{n}\left\|Q_{c(i)}-Q_{o(i)}\right\|$	0, ∞, 0	m^3/s	Legates & McCabe Junior (1999)
Mean absolute relative error (MARE)	$\frac{1}{n}\sum_{i=1}^{n}\frac{\left\|Q_{c(i)}-Q_{o(i)}\right\|}{Q_{o(i)}}$	0, ∞, 0	-	Rientjes et al. (2013)
Mean square error (MSE)	$\frac{1}{n}\sum_{i=1}^{n}\left(Q_{c(i)}-Q_{o(i)}\right)^{2}$	0, ∞, 0	$(m^3/s)^2$	Legates & McCabe Junior (1999)
Root mean square error (RMSE)	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Q_{c(i)}-Q_{o(i)}\right)^{2}}$	0, ∞, 0	m^3/s	Romanowicz et al. (2013)
Transformed root mean square error (TRMSE)	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{Q}_{c(i)}-\hat{Q}_{o(i)}\right)^{2}}$	0, ∞, 0	m^3/s	Kollat et al. (2012)
Ratio of RMSE to standard deviation of observations (RSR)	$\sqrt{\sum_{i=1}^{n}\left(Q_{c(i)}-Q_{o(i)}\right)^{2}}/\sqrt{\sum_{i=1}^{n}\left(Q_{o(i)}-\overline{Q_{o}}\right)^{2}}$	0, ∞, 0	-	Muleta (2012)
Modification of RMSE to high flow errors (NHF)	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Q_{c(i)}-Q_{o(i)}\right)^{2}*\left(\frac{Q_{o(i)}}{Q_{o(max)}}\right)^{2}}$	0, ∞, 0	m^3/s	Fenícia et al. (2007)

Legend: Qo(i) and Qc(i) are the observed and calculated daily streamflow at time-interval i, $\bar{Q}_{o(i)}$ and $\bar{Q}_{c(i)}$ are the observed and calculated mean streamflows, mQo(i) and mQc(i) are the total monthly streamflows observed and calculated, $\overline{Q_D}$ is the interannual calendar day mean observed streamflows, $Q_{ref(i)}$ is the average reference streamflows, $\hat{Q}_{o(i)}$ and $\hat{Q}_{c(i)}$ are the observed and calculated transformed streamflows, *K* is the total number of years in the time-series, ω is a weighting parameter (used ω = 0.1), $QD_c(p)$ and $QD_o(p)$ are the observed and calculated streamflow for the probability p of the duration curve, Δp is the interval used from the duration curve for the sum, and n is the total number of records in the time-series.

Table 1. Continued...

Name (Symbology)	Mathematical formulation	Min, Max, Optimal	Units	References
Modification of RMSE to low flow errors (NLF)	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Q_{c(i)}-Q_{o(i)}\right)^{2}*\left(\frac{Q_{o(max)}-Q_{o(i)}}{Q_{o(max)}}\right)^{2}}$	0, ∞, 0	m^3/s	Fenícia et al. (2007)
Sum of squared erros of the streamflows logarithmic (SLOGQ)	$\sum_{i=1}^{n}\left(logQ_{c(i)}-logQ_{o(i)}\right)^{2}$	0, ∞, 0	$(m^3/s)^2$	Hogue et al. (2000)
Sum squared errors of daily streamflows (SSEQ)	$\sum_{i=1}^{n}\left(Q_{c(i)}-Q_{o(i)}\right)^{2}$	0, ∞, 0	$(m^3/s)^2$	Wohling et al. (2013)
Sum squared errors of monthly streamflows normalized by basin area (SSEMQ)	$\frac{\sum_{i=1}^{n}\left(mQ_{c(j)}-mQ_{o(j)}\right)^{2}}{A}$	0, ∞, 0	$(m^3/s)^2/m^2$	Wohling et al. (2013)
Maximal absolute error (MAXAE)	$max\left\|Q_{c(i)}-Q_{o(i)}\right\|$	0, ∞, 0	m^3/s	Wohling et al. (2013)
Maximum difference in the largest peak flows (DHQMAX)	$max\left(Q_{c(i)}\right)-max\left(Q_{o(i)}\right)$	- ∞, ∞, 0	m^3/s	Wohling et al. (2013)
Relative volume error (ΔV)	$\left(\sum_{i=1}^{n}\left(Q_{c(i)}\right)-\sum_{i=1}^{n}\left(Q_{o(i)}\right)\right)/\sum_{i=1}^{n}\left(Q_{o(i)}\right)$	- ∞, ∞, 0	-	Rientjes et al. (2013)
Volumetric efficiency (VE)	$1-\left(\sum_{i=1}^{n}\left\|Q_{c(i)}-Q_{o(i)}\right\|/\sum_{i=1}^{n}Q_{o(i)}\right)$	-∞, 1, 1	-	Criss & Winston (2008)
Runnoff coefficient percent error (ROCE)	$\frac{1}{K}\sum_{k=1}^{K}\left\|\left(\bar{Q}_{c(year)}/\bar{Q}_{o(year)}\right)-1\right\|*100\%$	0, ∞, 0	-	Kollat et al. (2012)
Combined form of NSE and ΔV (Y)	$NSE/\left(1+\left\|\Delta V\right\|\right)$	-∞, 1, 1	-	Rientjes et al. (2013)
Combined form of NSE and MARE (RV)	$NSE-\omega*\left\|MARE\right\|$	-∞, 1, 1	-	Romanowicz et al. (2013)
Slope of the streamflow duration curve (SFDCE)	$\left\|\left(Q_{c,50\%}-Q_{c,10\%}/Q_{o,50\%}-Q_{o,10\%}\right)-1\right\|*100\%$	0, ∞, 0	-	Kollat et al. (2012)
Streamflow duration curve index (SDCI)	$1-\left[\sum_{p=po}^{np}\left(QD_c(p)-QD_o(p)\right)*\Delta p/\left(\sum_{p=po}^{np}QD_o(p)*\Delta p\right)\right]$	0, ∞, 1	-	Tucci (2005)

Legend: Qo(i) and Qc(i) are the observed and calculated daily streamflow at time-interval i, $\bar{Q}_{o(i)}$ and $\bar{Q}_{c(i)}$ are the observed and calculated mean streamflows, mQo(i) and mQc(i) are the total monthly streamflows observed and calculated, $\overline{Q_D}$ is the interannual calendar day mean observed streamflows, $Q_{ref(i)}$ is the average reference streamflows, $\hat{Q}_{o(i)}$ and $\hat{Q}_{c(i)}$ are the observed and calculated transformed streamflows, K is the total number of years in the time-series, ω is a weighting parameter (used ω = 0.1), $QD_c(p)$ and $QD_o(p)$ are the observed and calculated streamflow for the probability p of the duration curve, Δp is the interval used from the duration curve for the sum, and n is the total number of records in the time-series.

against a baseline model represented by the mean of all streamflow observations. An adaptation of NSE is the NSE calculated with the logarithm of the daily streamflows (LNS). In this way the oversensitivity of NSE to extreme values is reduced and the sensitivity for lower values is increased (Krause et al., 2005). Another adaptation of NSE with this aim is the modified form of the NSE (MNS), computed with the absolute value of the linear difference between *Qo* and *Qc* (Krause et al., 2005).

Other modifications to NSE are related to alternative benchmark models used instead of the mean of all streamflow observations (e.g. Schaefli & Gupta, 2007; Krause et al., 2005; Muleta, 2012). One of these metrics measures HM performance relative to a reference model given by the interannual calendar day mean (named NSD by Muleta, 2012). Similar to LNS, the LNSD (NSE with calendar day mean calculated on log transformed daily streamflows) was proposed. Following MNS, the MNSD (Modified form of NSE with calendar day mean) uses the absolute value of the linear differences. The NSE that use calendar monthly mean streamflow as a reference model (NSM) was used for daily HM calibration. Also derived from NSE, Persistence Index (PI) uses previously observed values as the reference model, which is appropriate in a streamflow forecasting context (Bennett et al., 2013). This index measures the relative magnitude of the residual variance against the variance of errors obtained by a persistence model (Gupta et al., 1999).

HF (high flow) metric was proposed to evaluate the performance of a HM in reproducing peak streamflow values (Rwetabula et al., 2012). Willmott's index of agreement (D) was proposed to overcome the limitation of r^2 related to poor HMs that consistently overestimate or underestimate the observations (Muleta, 2012). Another metric that resembles NSE is the Kling-Gupta Efficiency (KGE). This is an adaptation and at the same time decomposition of NSE, which facilitates the analysis of the relative importance of its different components (correlation, bias, and variability measure - α) in the context of HM calibration (Gupta et al., 2009). According to Pechlivanidis et al. (2012), the KGE sees the calibration process from a multi-objective optimization perspective. A modification of KGE has also been proposed by Pool et al. (2018) aiming at achievieng a non-parametric calibration criteria.

The normalized bias of flows (β) indicates the relationship between the mean flow difference (*Qo* and *Qc*) normalized by the standard deviation of the observed flows (Wohling et al., 2013).

In contrast to the metrics that follow NSE-like formulations, there are metrics based on the direct difference between *Qo* and *Qc*, which are therefore referred to as a type of error. Examples of this group of metrics are mean error (ME), mean absolute error (MAE), mean absolute relative error (MARE), mean square error (MSE), root mean square error (RMSE), and transformed root mean square error (TRMSE). ME is the average of the time-series of errors, thus it identifies whether the HM is more biased to overestimate or underestimate streamflows. However, it does not quantify these errors distinctly. Despite other metrics mentioned in this group do not compensate for the positive and negative error values like ME, their values do not indicate if the HM overestimates or underestimates the observations. MAE quantifies the average of the time-series of absolute values of the errors, while MARE quantifies the average of a time-series of absolute values of the error relative to the observed streamflow. MSE averages the time-series of squared errors, avoiding the error compensation of ME, but making the interpretation of the metric's value difficult as it is in a different unit (i.e. square m^3/s). RMSE overcomes the limitation of MSE by applying the root over MSE. TRMSE uses a Box-Cox transformation of the streamflow to quantify the RMSE. The Box-Cox transformation, in addition to emphasizing low-flow periods, also reduces the impact of heteroscedasticity in the RMSE calculation (Hogue et al., 2000; Kollat et al., 2012).

Other metrics are derivations of RMSE as RSR (ratio between RMSE and the standard deviation of the streamflow observations (Moriasi et al., 2007), NHF (modification of RMSE for increasing sensitivity to high-flow errors) and NLF (modification of RMSE for increasing sensitivity to low-flow errors) presented by Fenícia et al. (2007).

SLOGQ (sum of squared errors of the streamflows logarithm) metric is a function selected for the calibration of parameters that influence the hydrograph recessions (Hogue et al., 2000). SSEQ (Sum of squared errors of daily streamflows) and SSEMQ (Sum of squared errors of monthly streamflows normalized by basin area) metrics, although not calculating averages, have similarities to MSE because they represent the sum of squared deviations and result in distinct units of the variable under analysis, which makes interpretation difficult (Wohling et al., 2013).

The discrepancy between peak flow values is quantified by the MAXAE (maximal absolute error), which has the disadvantage of being subject to a time-interval error (Janssen & Heuberger, 1995). Metric DHQMAX (maximum difference in the largest peak flows) uses a timeless relationship to quantify the difference between maximum observed and calculated streamflows. Both metrics are directly related to errors in peak streamflows (Wohling et al., 2013).

ΔV (relative volume error) is usually called Bias and is the mean error between observed and calculated streamflows expressed as a fraction of the average observed streamflows (Rwetabula et al., 2012). It is commonly recommended for quantifying water balance errors (Rientjes et al., 2013) and indicates whether the model is poor in representativeness (Moriasi et al., 2007; Van Liew et al., 2007). VE (volumetric efficiency), on the other hand, evaluates the deviation between observed and calculated hydrographs by measuring the area between them, expressed as a fraction of the average observed streamflows (Criss & Winston, 2008). ROCE (runoff coefficient percent error) metric considers water balance as the average annual runoff coefficient percent error. As presented in Table 1, the sum occurs during years 1 to *k* of the calibration period, for which an average annual value is then calculated (Kollat et al., 2012).

Other metrics combine previously presented metrics to measure more than one issue, as Y (combined form of NSE and ΔV (Akhtar et al., 2009)) and RV(combined form of NSE and MARE (Lindström et al., 1997), weighted by a parameter ω. The best results of this metric are obtained with ω equals to 0.1 according to the application of the HBV hydrologic model by Lindström et al. (1997) and Dakhlaoui et al. (2012).

Finally, SFDCE (slope of the streamflow duration curve) and SDCI (streamflow duration curve index) metrics refer to the comparison between the calculated and observed streamflow duration curves. SFDCE represents the error in simulating the slope of the streamflow duration curve (Westerberg et al., 2011; Kollat et al., 2012). SDCI evaluates the similarity between the observed and calculated streamflow duration curves from the sum of the differences between all the points that define the curves (Tucci, 2005).

METHODOLOGY

The metrics showed in Table 1 were applied to assess the performance of the calculated and synthetic streamflow time-series in the Piancó River and Furnas subcatchments relative to the observed streamflows. This procedure aims to evaluate how metrics are influenced by the quality of the synthetic streamflow time-series, and also to compare metrics from synthetic time-series to metrics from a calculated streamflow time-series obtained from a calibrated HM. Based on the results of this procedure, a critical analysis of each metric was carried out, showing use recommendations and limitations.

A four-step procedure was used and is described in the following sections: 1) metrics selection, 2) data collection from two case studies, 3) definition of ten synthetic streamflow time-series, and 4) test analysis and results.

Metrics selection

A total of 36 metrics commonly used as OFs for HM calibration were identified and selected from an extensive literature review, considering issues related to their frequency of use; whether they are modifications, adaptations or combinations of preexistent metrics; or comprise new concepts, as explained in the previous section and summarized in Table 1. It is worth mentioning that we have not used these metrics for model calibration, but rather to provide the evaluation of the output of a hydrologic model previously calibrated and also of hypothetical cases based on synthetic time series, as further detailed.

Case studies

Two case studies were selected based on data availability and the existence of previously calibrated HMs, with distinct hydrological regimes (intermittent or perennial rivers) and drainage area, in order to provide a broader picture regarding the results and findings.

The first case study is a subcatchment of the Piancó (drainage area of 4,603.39 km^2), located in the Piancó River basin (Figure 1B) in northeast Brazil. This is a semiarid region with a large number of intermittent rivers. This study used daily time-series of observed streamflow from 1970 to 2011 (42 years) from Felix & Paz (2016), in which the MGB-IPH model (Collischonn et al., 2007) was applied to the subcatchment.

The hydrological regime of the Piancó River is characterized by strong seasonality, with monthly streamflow ranging from 405.49 to 0.09 m^3/s in the rainy season (January to May) and typically zero flows in the driest moths. The river was dry in 37% of the daily time intervals of the time-series and the driest year was 1980, which saw 79% of the days without streamflow. The MGB-IPH model was calibrated and validated by Felix & Paz (2016) for the periods 1970-1990 and 1991-2011, respectively, through an automatic multi-objective calibration procedure (using NSE, LNS and ΔV as OFs), followed by a calibration refinement procedure done manually in order to obtain more representative and coherent parameters between different hydrological response units. A total of 11 parameters was calibrated as detailed in the mentioned reference.

The second case study covers the Furnas subcatchment with a drainage area of 51,784.41 km^2, located in the Grande River Basin in southeastern Brazil, in the Paraná hydrographic region (Figure 1C). This basin is widely used for hydroelectric power generation (Tucci et al., 2008). This subcatchment was modeled by Bravo et al. (2009), who applied the MGB-IPH model

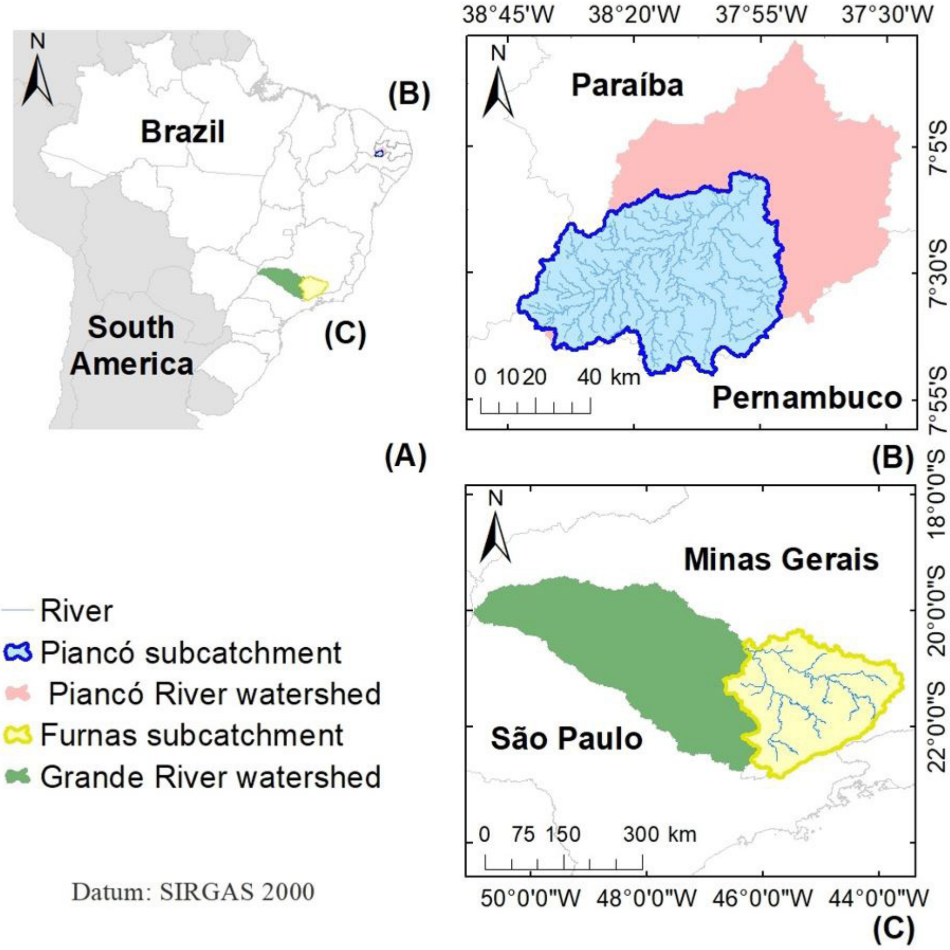

Figure 1. Location of study areas: (A) in Brazil; (B) Piancó subcatchment and (C) Furnas subcatchment.

using daily streamflow data from 1981 to 2001 (21 years). The hydrological regime in Furnas is strongly seasonal, ranging from 350 m^3/s during low flows to over 2,000 m^3/s in summer, with flood peaks typically reaching 4,000 m^3/s (Bravo et al., 2009). These authors applied the MGB-IPH for the period 1970-1980 during calibration. Validation was carried out for the period 1981-2001. As with the first study case, calibration of the MGB-IPH to Furnas subcatchment was performed through an automatic multi-objective calibration procedure with the same OFs, considering a total of 10 parameters, as detailed in the mentioned reference.

For both study cases, it was used the version of the MGB-IPH model that adopts a square-grid discretization, as presented in Collischonn et al. (2007). Pianco subcatchment was divided in 151 cells of approximately 5 x 5 km and Furnas subcatchment was discretized into 519 cells of roughly 10 x 10 km. This model was selected due to satisfactory results being achieved on several applications in different hydrologic regimes (e.g. Oliveira et al., 2018; Pereira et al., 2014; Paiva et al., 2013; Ribeiro Neto et al., 2006; Tucci et al., 2005) and due to availability of previous works by the authors. But, in fact, this study could have been performed considering the outputs of any calibrated hydrologic model.

Synthetic streamflow time-series

Eleven daily streamflow time-series were used in the analysis carried out in each watershed.

One of these time-series used the daily calculated streamflow (Qhid) from the previous studies of Felix & Paz (2016) for the Piancó subcatchment, and Bravo et al. (2009) for the Furnas subcatchment. The Qhid time series were useful for providing the basis for developing the synthetic time-series and also for serving as comparison to these time-series, as detailed bellow.

Ten synthetic daily streamflow time-series were generated based on the calculated and observed values in each watershed, as result of idealized error behavior in hypothetical cases (Figure 2). The general idea is simple and of practical understanding: to analyze how does each metric evaluate hypothetical cases that present isolated very well known error behaviors. We want to assess if the metric is able to detect this known error or if the metric considered it as a perfect model; if there is a compensation effect between systematic errors and perfect match in distinct time periods; how much do the metrics penalize each type of well known error or valorized each type of perfect model capability; and how the evulation of these hypothetical cases relatively to an actual typical output of a calibrated hydrologic model. These synthetic time series represent in some cases exaggerated systhematic errors or model capabilities that do occur when calibrating a hydrological model but at smaller intensity and not isolated from other errors.

For example, the synthetic time-series Qox2 (Qo/2) shows in each time interval a streamflow value that is equal to twice (half) that which was observed. These time series were proposed to detect how each metric evaluated an hypothetical case that systematically calculates half or double of the discharges in each time step. They are perfect models in terms of predicting timing of recession and peak flows, for instance. And also we intended to analyse if each metric evaluated the actual calibrated model better or worse than these Qox2 and Qo/2 hypothetical cases.

Two other synthetic time-series were based on the use of the Q50 (median of the observed streamflow time-series), which was equal to 0.23 m^3/s in the Piancó subcatchment and to 703 m^3/s in the Furnas subcatchment. Thus, the synthetic time-series Qo+Q50 (Qo-Q50) shows a streamflow value that is equal to the observed one plus (minus) Q50 in each time interval. If the resulting value of the streamflow for Qo-Q50 was less than zero in a time interval, it was considered as zero. The Qo+Q50 and Qo-Q50 time series represent hypothetical cases that systematically shift up or down, respectively, the observed hydrograph by a constant value.

Two synthetic time-series combine calculated and observed streamflows over different time periods within the year. The synthetic time-series Qo_{wet} (Qo_{dry}) shows observed streamflow values in the wet (dry) periods and calculated streamflow values in the dry (wet) periods of each year. The Qo_{wet} time series

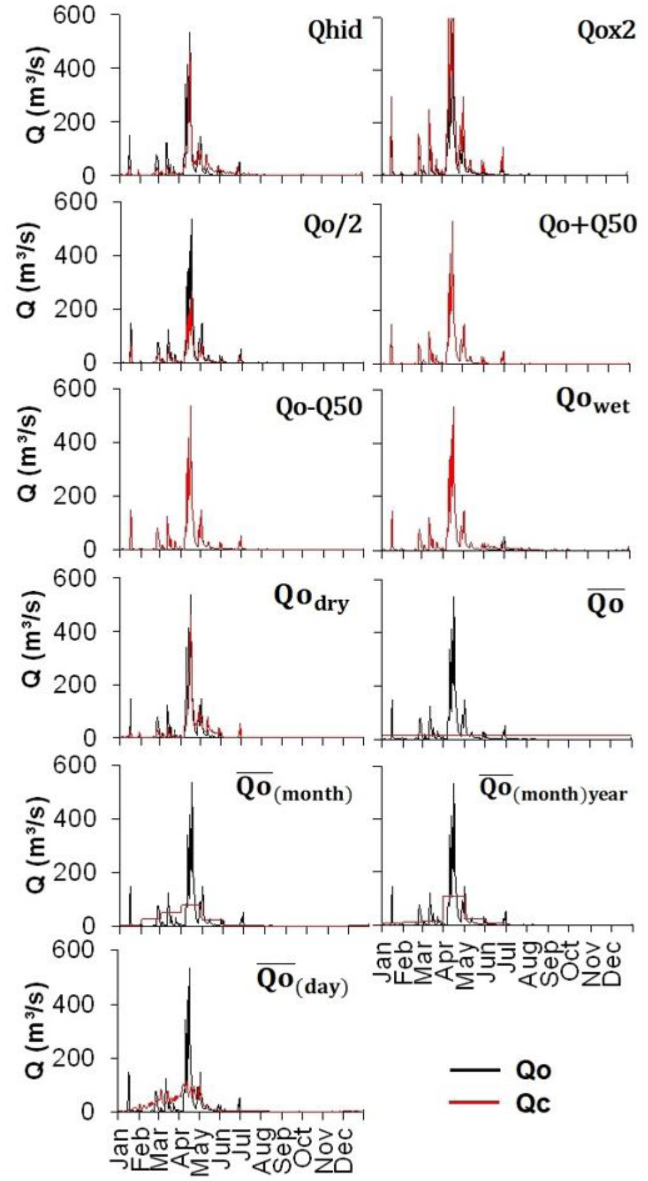

Figure 2. Observed, calculated and synthetic daily streamflows in 1973, Piancó subcatchment.

represents a hypothetical case that is perfect during wet period in reproducing observed values, while maintaining the typical error of a calibrated model during the dry period. Analogously, the Qo_{dry} time series is like a hypothetical case in which it perfectly reproduces observed flows during the dry period and presents typical error during the wet period.

The last four synthetic time-series were based on mean values derived from the observed streamflow time-series. The idea is to have hypothetical cases that conservatively predict streamflow following the historic pattern according to the mean values at different ways. The synthetic time-series $\overline{Qo}$ is simply the same streamflow value in each day, equal to the mean observed streamflow. The synthetic time-series $\overline{Qo}_{(month)}$ shows the same streamflow value in each day of a given month, equal to the mean observed streamflow derived with data of that month in all years of the observed time-series. Thus, this time-series is comprised of 12 distinct values, repeated for every year. The synthetic time-series $\overline{Qo}_{(month)year}$ shows the same streamflow value in each day of a given month, equal to the mean observed streamflow of that specific month. In this way, the streamflow values are distinct between months in a given year and in another year. Finally, any daily streamflow in the synthetic time-series $\overline{Qo}_{(day)}$ is equal to the mean observed streamflow derived from the data for that day in all years of the observed time-series. Between years, the daily streamflow values are the same in each day.

RESULTS

Performance of the synthetic streamflow time-series

Results of the performance assessment of calculated and synthetic time-series by the 36 selected metrics are discussed below (Table 2 and Figure 3).

Table 2. Performance of the synthetic streamflows time-series related to the performance of calculated streamflows (Qhid) (F and P indicate Furnas and Piancó subcatchments, respectively; green upward arrow: higher metric value; red downward arrow: lower metric value; circle: exactly same metric value).

OFs	Units	Qhid		Qox2		Qo/2		Qo+Q50		Qo-Q50		Qo_{wet}		Qo_{dry}		$\overline{Qo}$		Qo_{month}		$Qo_{(month)yaer}$		Qo_{year}	
		F	P	F	P	F	P	F	P	F	P	F	P	F	P	F	P	F	P	F	P	F	P
r	-	0.95	0.85	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↓	↓	↓	↓	↓	↓	↓	↓
r²	-	0.90	0.72	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↓	↓	↓	↓	↓	↓	↓	↓
NSE	-	0.89	0.72	↓	↓	↓	↑	↓	↑	↓	↑	↑	↑	↑	↑	↓	↓	↓	↓	↓	↓	↓	↓
LNS	-	0.90	-0.34	↓	↑	↓	↑	↓	↑	↓	↑	↑	↑	↑	↑	↓	↓	↓	↓	↓	↑	↓	↓
MNS	-	0.71	0.61	↓	↓	↓	↑	↓	↑	↓	↑	↑	↑	↑	↑	↓	↓	↓	↓	↓	↓	↓	↓
NSD	-	0.80	0.66	↓	↓	↓	↑	↓	↑	↓	↑	↑	↑	↑	↑	↓	↓	↓	↓	↓	↓	↓	↓
LNSD	-	0.90	0.19	↓	↑	↓	↑	↓	↑	↓	↑	↑	↑	↑	↑	↓	↓	↓	↓	↓	↑	↓	↓
MNSD	-	0.54	0.46	↓	↓	↓	↑	↓	↑	↓	↑	↑	↑	↑	↑	↓	↓	↓	↓	↓	↓	↓	↓
NSM	-	0.49	0.51	↓	↓	↓	↑	↓	↑	↓	↑	↑	↑	↑	↑	↓	↓	↓	↓	↓	↓	↓	↓
PI	-	-4.53	0.14	↓	↓	↓	↑	↓	↑	↓	↑	↑	↑	↑	↑	↓	↓	↓	↓	↓	↓	↓	↓
HF	-	0.90	0.83	↓	↓	↓	↓	↓	↑	↓	↑	↑	↑	↑	↑	↓	↓	↓	↓	↓	↓	↓	↓
D	-	0.97	0.92	↓	↓	↓	↓	↓	↑	↓	↑	↑	↑	↑	↑	↓	↓	↓	↓	↓	↓	↓	↓
α	-	0.98	0.90	↓	↓	↓	↓	↑	↑	↓	↑	↑	↑	↑	↑	↓	↓	↓	↓	↓	↓	↓	↓
β	-	0.10	0.00	↓	↓	↓	↓	↓	↓	↓	↑	↑	↓	↑	↓	↑	↑	↑	↓	↑	↑	↑	↓
KGE	-	0.91	0.82	↓	↓	↓	↓	↓	↑	↓	↑	↑	↑	↑	↓	↓	↓	↓	↓	↓	↓	↓	↓
ME	(m³/s)	75.11	0.14	↓	↓	↓	↓	↓	↓	↓	↑	↑	↓	↑	↓	↑	↑	↑	↓	↑	↑	↑	↓
MAE	(m³/s)	161.28	10.01	↓	↓	↓	↑	↓	↑	↓	↑	↑	↑	↑	↓	↓	↓	↓	↓	↓	↓	↓	↓
MARE	-	18.89	$3{\cdot}10^6$	↓	↑	↓	↑	↓	↑	↓	↑	↑	↑	↑	↑	↓	↓	↓	↓	↓	↓	↓	↓
MSE	(m³/s)²	63672	1153	↓	↓	↓	↑	↓	↑	↓	↑	↑	↑	↑	↑	↓	↓	↓	↓	↓	↓	↓	↓
RMSE	(m³/s)	252.33	33.96	↓	↓	↓	↑	↓	↑	↓	↑	↑	↑	↑	↑	↓	↓	↓	↓	↓	↓	↓	↓
TRMSE	(m³/s)	1.56	1.58	↓	↑	↓	↑	↓	↑	↓	↑	↑	↑	↑	↑	↓	↓	↓	↓	↓	↓	↓	↓

Legend: Qhid is the time-series of calculated streamflows; F and P indicate Furnas and Piancó subcatchments, respectively. Linear correlation coefficient (r), Coefficient of determination (r²), Nash-Sutcliffe efficiency (NSE), NSE on log transformed daily flows (LNS), Modified forms of NSE (MNS), NSE with calendar day mean (NSD), NSE with calendar day mean calculated on log transformed daily flows (LNSD), Modified form of NSE with calendar day mean (MNSD), NSE with calendar monthly mean as reference model (NSM), Persistence Index (PI), High flow (HF), Index of agreement (D), Relative variability (α), Normalised bias of flows (β), Kling-Gupta efficiency (KGE), Mean error (ME), Mean absolute error (MAE), Mean absolute relative error (MARE), Mean square error (MSE), Root mean square error (RMSE), Transformed root mean square error (TRMSE), Ratio of RMSE to standard deviation of observations (RSR), Modification of RMSE to high flow errors (NHF), Modification of RMSE to low flow errors (NLF), Sum of squared erros of the streamflows logarithmic (SLOGQ), Sum squared errors of daily streamflows (SSEQ), Sum squared errors of monthly streamflows normalized by basin area (SSEMQ), Maximal absolute error (MAXAE), Maximum difference in the largest peak flows (DHQMAX), Relative volume error (ΔV), Volumetric efficiency (VE), Runnoff coefficient percent error (ROCE), Combined form of NSE and ΔV (Y), Combined form of NSE and MARE (RV), Slope of the streamflow duration curve (SFDCE) and Streamflow duration curve index (SDCI).

Table 2. Continued...

OFs	Units	Qhid		Qox2		Qo/2		Qo+Q50		Qo-Q50		Qo_{wet}		Qo_{dry}		$\overline{Qo}$		Qo_{month}		$Qo_{(month)yaer}$		Qo_{year}	
		F	P	F	P	F	P	F	P	F	P	F	P	F	P	F	P	F	P	F	P	F	P
RSR	-	0.33	0.53	↓	↓	↓	↑	↓	↑	↓	↑	↑	↑	↑	↑	↓	↓	↓	↓	↓	↓	↓	↓
NHF	(m^3/s)	96.30	10.19	↓	↓	↓	↓	↓	↑	↓	↑	↑	↑	↑	↑	↓	↓	↓	↓	↓	↓	↓	↓
NLF	(m^3/s)	177.29	27.01	↓	↓	↓	↑	↓	↑	↓	↑	↑	↑	↑	↑	↓	↓	↓	↓	↓	↓	↓	↓
SLOGQ	(m^3/s)	59.93	$2{\cdot}10^5$	↓	↑	↓	↑	↓	↑	↓	↑	↑	↑	↑	↑	↓	↓	↓	↓	↓	↑	↓	↓
SSEQ	$(m^3/s)^2/m^2$	$5{\cdot}10^8$	$8{\cdot}10^6$	↓	↓	↓	↑	↓	↑	↓	↑	↑	↑	↑	↑	↓	↓	↓	↓	↓	↓	↓	↓
SSEMQ	$(m^3/s)^2$	0.19	0.01	↓	↓	↓	↓	↓	↑	●	↑	↑	↑	↑	↑	↓	↓	↓	↓	↑	↑	↓	↓
MAXAE	(m^3/s)	5028.1	543.2	↓	↓	↑	↓	↑	↑	↑	↑	↑	↑	●	●	↓	↓	↓	↓	↑	↓	↓	↓
DHQMAX	(m^3/s)	1826.7	-324	↓	↓	↓	↓	↑	↑	↑	↑	↑	↑	●	●	↓	↓	↓	↓	↓	↓	↓	↓
ΔV	%	7.72	0.86	↓	↓	↓	↓	↓	↓	↓	↑	↑	↓	↑	↓	↑	↑	↑	↓	↑	↑	↑	↓
VE	-	0.83	0.39	↓	↓	↓	↑	↓	↑	↓	↑	↑	↑	↑	↑	↓	↓	↓	↓	↓	↓	↓	↓
ROCE	%	13.14	35.20	↓	↓	↓	↓	↓	↑	↓	↑	↑	↑	↑	↑	↓	↓	↓	↓	↑	↑	↓	↓
Y	-	0.83	0.71	↓	↓	↓	↓	↓	↑	↓	↑	↑	↑	↑	↓	↓	↓	↓	↓	↓	↓	↓	↓
RV	-	0.87	$- 3{\cdot}10^6$	↓	↑	↓	↑	↓	↑	↓	↑	↑	↑	↑	↑	↓	↓	↓	↓	↓	↓	↓	↓
SFDCE	%	8.84	13.12	↓	↓	↓	↓	↑	↓	↑	↓	↑	↑	↑	↑	↓	↓	↓	↓	↑	↓	↓	↓
SDCI	-	0.93	1.05	↓	↓	↓	↓	↓	↓	↓	↓	↑	↑	↑	↓	↑	↓	↑	↓	↑	↓	↑	↓

Legend: Qhid is the time-series of calculated streamflows; F and P indicate Furnas and Piancó subcatchments, respectively. Linear correlation coefficient (r), Coefficient of determination (r^2), Nash-Sutcliffe efficiency (NSE), NSE on log transformed daily flows (LNS), Modified forms of NSE (MNS), NSE with calendar day mean (NSD), NSE with calendar day mean calculated on log transformed daily flows (LNSD), Modified form of NSE with calendar day mean (MNSD), NSE with calendar monthly mean as reference model (NSM), Persistence Index (PI), High flow (HF), Index of agreement (D), Relative variability (α), Normalised bias of flows (β), Kling-Gupta efficiency (KGE), Mean error (ME), Mean absolute error (MAE), Mean absolute relative error (MARE), Mean square error (MSE), Root mean square error (RMSE), Transformed root mean square error (TRMSE), Ratio of RMSE to standard deviation of observations (RSR), Modification of RMSE to high flow errors (NHF), Modification of RMSE to low flow errors (NLF), Sum of squared erros of the streamflows logarithmic (SLOGQ), Sum squared errors of daily streamflows (SSEQ), Sum squared errors of monthly streamflows normalized by basin area (SSEMQ), Maximal absolute error (MAXAE), Maximum difference in the largest peak flows (DHQMAX), Relative volume error (ΔV), Volumetric efficiency (VE), Runnoff coefficient percent error (ROCE), Combined form of NSE and ΔV (Y), Combined form of NSE and MARE (RV), Slope of the streamflow duration curve (SFDCE) and Streamflow duration curve index (SDCI).

The synthetic streamflow time-series Qox2 represents the output of a hypothetical case that always doubled the observed values. Such a time-series presents perfect linear correlation with Qo and, therefore, the r and r^2 metrics reached the maximum value, superior to the Qhid performance for both basins, as expected. For the Furnas subcatchment, all other metrics assessed Qox2 performance as inferior to the ones obtained with Qhid. Due to intermittence and very low streamflows in the Piancó subcatchment, however, metrics that use logarithm of streamflows (e.g. LNS, LNSD and SLOGQ) assessed Qox2 performance as much better than Qhid, which showed difficulty in representing low streamflow values(Figures 3D, 3G and 3W). Furthermore, the TRMSE metric assessed Qox2 performance higher than Qhid, as this metric uses a transformation of the streamflows that expands the lower end of the scale and thus gives higher emphasis to recessions (Figure 3U). All other metrics assessed Qox2 performance as inferior to the ones obtained with Qhid in the Piancó subcatchment (Table 2).

The Qox2 performance in both subcatchments was lower when assessed by NSE, NSD, NSM, PI, HF, D KGE, and RSR metrics, which use the square of the residual in their formulation. Similar lower performance results were obtained by error-type metrics, whether they compute squared, absolute, or linear errors, as with ME, MAE, MARE, MSE, RMSE, NHF, NLF, SSEMQ, SSEQ, ROCE, or DHQMAX and MAXAE. These metrics are sensitive to systematic overestimation of streamflows, especially during floods, whether or not the river is intermittent. The MARE metric shows higher values in most of the synthetic time-series when compared to the MARE obtained with Qhid. Its values were very high for the Piancó River subcatchment (Figure 3R) due to recurrent zero streamflows. This factor also was reflected in RV values, as this is a MARE-dependent metric.

The synthetic streamflow time-series Qo/2 is similar to Qox2 and represents a streamflow value that is equal to half the observed one, in each time interval. For this reason, the r and r^2 metrics had the maximum value for Qo/2 in both subcatchments, as expected (Figure 3A and 3B). For Furnas, the MAXAE metric showed a lower value for Qo/2 than for Qhid, meaning the HM outputs are lower than half of the Qhid values in some time intervals during flood periods (Table 2). For the Piancó River subcatchment, the performance results for Qo/2 were quite different from the results from the Furnas subcatchment, except for r and r^2. The performance of the Qo/2 time-series was assessed as better than the Qhid performance by more than half of the metrics. Among the metrics that did not follow this behavior are ME, ΔV, KGE and metrics based on streamflow duration curves as SFDCE and SDCI (Figures 3P, 3AD, 3O, 3AI, and 3AJ). These latter metrics did not perform satisfactorily for both subcatchments.

The performance of two synthetic time-series that increased (Qo+Q50) or decreased (Qo-Q50) by a constant quantity (Q50) the observed streamflow values was assessed. Both time-series

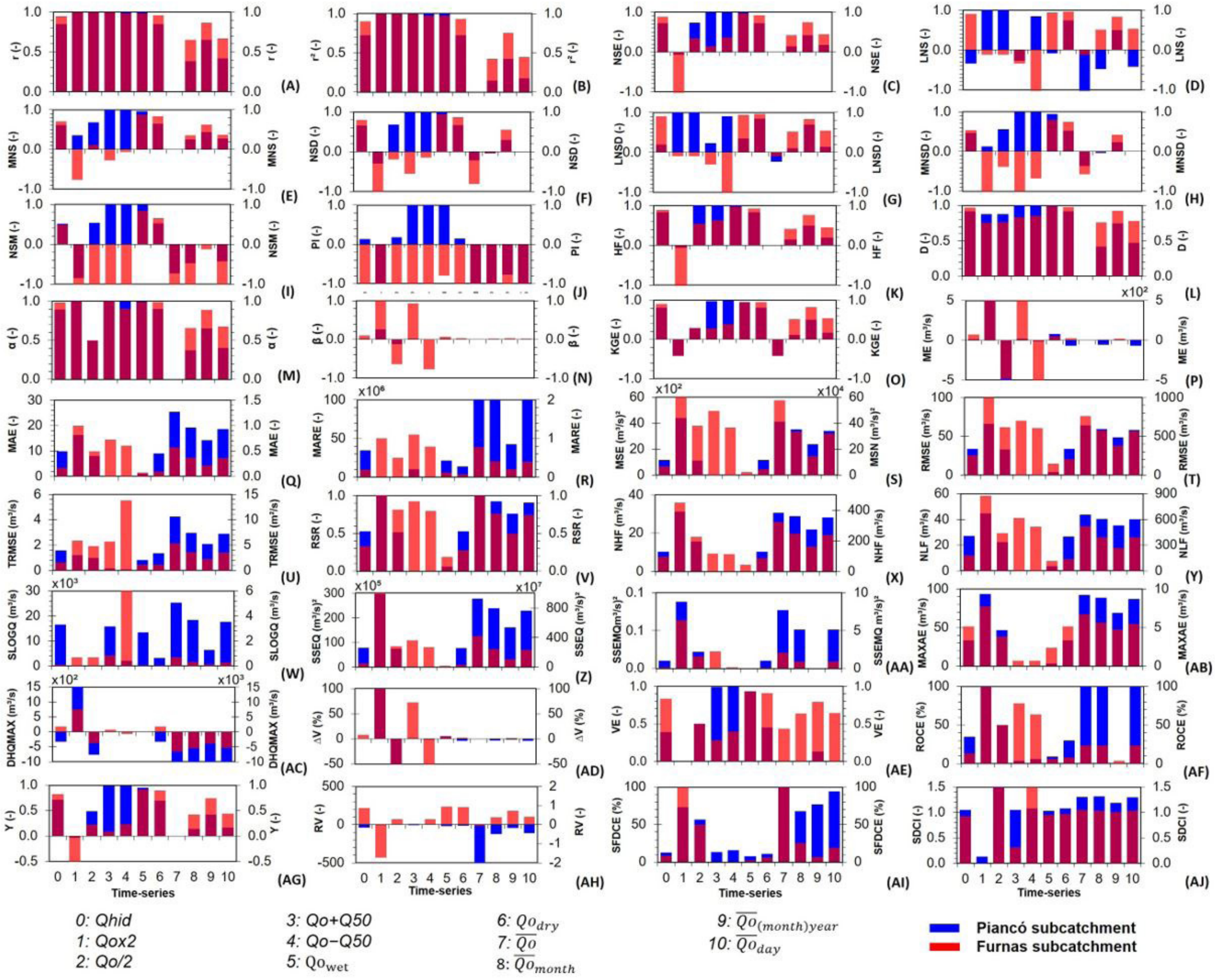

Figure 3. Performance metrics values for several streamflow time-series of a single catchment: results for Piancó river and Furnas subcatchments.
Legend: Qhid is the time-series of calculated streamflows.

showed better performance than Qhid in the Furnas subcatchment when assessed by r, r², MAXAE, DHQMAX, and SFDCE, while α showed a better performance only for Qo+Q50. The result for r and r² are the same as the other time-series that present a linear correlation with Qo. For the other mentioned metrics, even with a high Q50 value in the Furnas subcatchment (Q50 = 703 m³/s, which is 72% of the average daily flow and 9.4% of the maximum daily flow), this was not enough to cause peak streamflow errors (which are the focus of MAXAE and DHQMAX) greater than those in Qhid time-series. The performance of both time-series was optimal when assessed by the SFDCE metric since the slope of the streamflow duration curves is exactly the same as the observed one (Figure 3AI).

In the case of the Piancó River subcatchment, Q50 is extremely low (Q50 = 0.23 m³/s, equivalent to 1.4% of the daily average streamflow and < 0.02% of the daily maximum), making the Qo+Q50 and Qo-Q50 time-series very similar to Qo. Thus, most of the metrics showed superior performance on these synthetic time-series when compared to Qhid, except for the ME, ΔV, and β, while SFDCE and SDCI showed superior performance only for Qo-Q50. The slightly superior performance of Qo-Q50 compared to Qo+Q50 was because of the occurrence of zero streamflow values, meaning that for these days Qo-Q50 = Qo. The effect of having streamflows equal to zero in the observed time-series is also responsible for the slope of the duration curve of these synthetic time-series being different from the observed one. Thus, SFDCE metrics did not reach the ideal value in this subcatchment, as occurred for Furnas subcatchment.

The errors in the low streamflows in the synthetic time-series Qo/2 and Qo-Q50 in the Piancó River subcatchment did not affect the performance assessed by the LNS, LNSD, LOGQ, and TRMSE metrics (Figures 3D, 3G, 3W, and 3U). Unlike in the Piancó River subcatchment, the Qo-Q50 synthetic time-series for the Furnas subcatchment present higher errors, reducing its performance when assessed by those metrics.

The synthetic streamflow time-series Qo_{wet} and Qo_{dry} represent the output of a hypothetical case that has no error in the wet (dry) periods, and keeps the error of the adjusted HM in the opposite period. Thus, Qo_{wet} and Qo_{dry} should be evaluated as having a better performance than Qhid. Actually, the performance of both time-series was assessed as better than the performance of Qhid in almost all metrics. Only two metrics assessed Qo_{dry} with same performance as Qhid: MAXAE and DHQMAX (Table 2). As these two metrics assess the largest error during floods (wet period), the maximum flood error found in Qo_{dry} was equal to the one in Qhid (Figures 3AB and 3AC).

Several metrics which compensate for positive and negative errors assessed Qo_{wet} and Qo_{dry} as having a lower performance than Qhid in the Piancó River subcatchment. This highlights the negative aspect of such metrics (e.g. ME, ΔV and β), as errors of overestimation or underestimation are compensated for. That means that a hypothetical case that reproduces exactly the wet season but has errors during the dry period, when compared to another hypothetical case that also shows errors in the wet period, will present lower performance when assessed by these metrics (the same results would occur when changing the wet/dry periods). For example, the ME of Qhid in the Piancó River subcatchment was -0.14 m^3/s, while the ME of Qo_{wet} was -0.80 m^3/s and the ME of Qo_{dry} was 0.66 m^3/s (note that the sum of the latter two MEs are equal to the ME of Qhid). Since the metric Y uses ΔV and NSE (which do not compensate errors) in its formulation, this effect was not predominant, but led to Y assessing Qo_{dry} as of lower performance than Qhid. It is important to emphasize that this result is local-dependent, as a distinct adjusted HM error behavior in wet and dry periods could occur (e.g. if just positive or negative errors occur in both periods, there will not be a compensation effect).

The remaining synthetic time-series present daily streamflow values that are based on temporal averages derived from the observed time-series: $\overline{Qo}$ (mean streamflow), $\overline{Qo}_{(month)}$ (mean monthly streamflow), $\overline{Qo}_{(month)year}$ (mean monthly streamflow by year), and $\overline{Qo}_{(day)}$ (mean daily streamflow).

For the Furnas subcatchment, which has perennial rivers, these four synthetic time-series were assessed as having lower performance than Qhid by most of the metrics (Table 2). Few metrics assessed the performance of these time-series as better than Qhid: the metrics with compensating errors effect (e.g. ME, ΔV, and β); DHQMAX, which focuses on a point error; and SDCI which evaluates the similarity between streamflow duration curves. Thus, as synthetic time-series are based on average values, errors in high and low values are compensated for, avoiding larger errors in higher streamflows. In addition, the time-series $\overline{Qo}_{(month)year}$ showed a better performance than Qhid when assessed by MAXAE, ROCE, and SFDCE, as this time-series present a lower error in the maximum daily streamflow, in the average annual runoff coefficient, and in the slope of the streamflow duration curve.

For the Piancó River subcatchment, the performance results of the four synthetic streamflow time-series, based on mean values of the observed streamflows, were partially the same as in the Furnas subcatchment. The error-compensating effect of the ME, ΔV, and β metrics improved the performance of $\overline{Qo}$ and $\overline{Qo}_{(month)year}$ time-series when compared to Qhid, as also ROCE metric. But a distinct pattern was observed in the Piancó River subcatchment in logarithm-based metrics (e.g. LNS, LNSD, and SLOGQ). These metrics assessed the performance of only $\overline{Qo}_{(month)year}$ as better than Qhid. This means that the burden of the HM errors in reproducing the streamflow in the dry period in the Piancó River subcatchment, with intermittent rivers, was large enough for metrics LNS, LNSD, and SLOGQ to assess the performance of Qhid as lower than a synthetic time-series with mean monthly streamflow by year. However, Qhid performance assessed by these metrics was higher than the performance of time-series based on mean streamflows, mean monthly streamflows, and even mean daily streamflows.

Closure to response of performance metrics

It is well described in literature that each metric used for hydrologic model calibration has been proposed focused on one or some aspects of the comparison between calculated and observed streamflows (e.g. Gupta et al., 1998; Wohling et al., 2013; Pushpalatha et al., 2012; Madsen, 2000). As evidenced by our results, systematic or large errors in other aspects non-focused by each metric may not be accounted or may not have significant effect in its evaluation. Users could, therefore, conduct a misjudgement of the overall behaviour of their model. For example, correlation coefficient and coefficient of determination evaluate the linear correlation of the data. A hypothetical time-series that systematically doubled the discharges is evaluated as perfect by those metrics, while a distributed model carefully calibrated using state-of-the art method does not achieve such performance, as expected. This is a classic example in literature, but there were other situations we found and that were more distinct from those previously discussed in literature.

For instance, it could be highlighted the hypothetical time-series that represent a perfect reproduction of observed flows during the dry or wet periods and present a behaviour exactly the same of the calibrated hydrological model in the opposite period. It means that these hypothetical cases are better or equal to the hydrological model throughout the year in reproducing observed flows. There is no doubt about that, it is conceptual. However, metrics that are practically restricted to assessing wet periods (Maximal absolute error; and Maximum difference in the largest peak flows) were not influenced whether the model was perfect or not during the dry period. More importantly, metrics that make compensation of positive and negative errors (e.g. mean error, relative volume error, combined form of NSE and ΔV and normalised bias of flows) may lead to the judgement that a model being wrong in both wet and dry periods may be of better performance than being wrong just in one of these periods (considering the same behaviour in the other time period).

The results obtained with hypothetical cases that reproduce temporal averages of discharges provided another interesting question: how useful is a calibrated HM that performs worse than simply assuming as model prediction the monthly or other average discharges on time based on observed time series for a past period of time? If we could simply construct such average discharges time series, why to spend time and effort in developing hydrological models that perform worse? But is the calibrated HM really worse than those hypothetical time-series? Two issues need to be discussed to think about the answers for all these questions.

First of all: a better or worse model for what? The purpose of the model, for which it will be used for, is crucial for properly answering the usefulness of each model. For instance, whether the model will be used to estimate and manage water resources availability in dry periods or to estimate flood impacts of climate change scenarios request distinct model capabilities as priority.

This discussion leads the question about 'how good is a model?' to move towards the second point, the issue we addressed within this study, regarding 'how good is a metric to evaluate a model?'. This second question is linked to the first one and concerns the way we evaluate model performance. The aim of the model use should be always in mind as a major driver for selecting metrics for model evaluation. In the first case, for a model being applied for managing water resources availability in dry periods, the reproduction of observed recession flows is crucial and thus model calibration should focus on this issue. Our results recommend the use of metrics such as NSD, KGE and RV for perennial rivers and LNSD, TRMSE and Y for intermittent ones. For the second case, the estimation of flood impacts using hydrological modeling, model calibration should give emphasis on adjusting peak flows. Metrics such as MAE, RSR, ΔV and SFDCE are recommended, independently of the river being intermittent or not.

CONCLUSION

This study assessed 36 metrics that are frequently used for HM calibration by comparing calculated and observed hydrographs. Daily streamflow time-series were used from calculated values by MGB-IPH model from previous studies and ten synthetic time-series generated based on the calculated and observed values, as a result of idealized error behavior in hypothetical cases. Two Brazilian large-scale watersheds with contrasting characteristics were adopted as case studies.

This study highlighted that knowing the limitations and recommendations of a metric used as an OF is important for adequately evaluating a HM output in terms of observed flow regime reproduction. It is already known that the parameter values obtained through the calibration process are influenced by the OF selected. As the calculated streamflows are dependent on the parameter values, this means the OF must be chosen according to the reason for the use of the HM. The purpose for which the model will be used for is decisive for properly answering the usefulness of each calibrated model.

Our results reassert that each metric should be interpreted specifically thinking about the aspects it has been proposed for. In this sense, simultaneously taking into account a set of metrics would lead to a broader evaluation of HM ability. This highligts to the advantages of adopting a multiobjective model evaluation by combining metrics that assess distinct aspects.

For this it is important to initially understand the actual behaviour of observed streamflows. This analysis should not be disregarded and will be crucial for adequately interpreting metrics results of HM evaluation.

This study supplies a guideline for the choice of OFs, while the use of synthetic time series as those proposed in this work could be useful as an auxiliary step towards better understanding the evaluation of a calibrated hydrological model for each study case.

ACKNOWLEDGEMENTS

The authors thank CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico - Brazil) for funding a research project to which this work is linked and for providing scholarships for the the 2nd and 3rd authors, and to Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brazil (CAPES) - Finance Code 001 for the first author's scholarship.

REFERENCES

Akhtar, M., Ahmad, N., & Booij, M. J. (2009). Use of regional climate model simulations as input for hydrological models for the Hindukush–Karakorum–Himalaya region. *Hydrology and Earth System Sciences*, *13*(7), 1075-1089. http://dx.doi.org/10.5194/hess-13-1075-2009.

Bennett, N. D., Croke, B. F. W., Guariso, G., Guillaume, J. H. A., Hamilton, S. H., Jakeman, A. J., Marsili-Libelli, S., Newham, L. T. H., Norton, J. P., Perrin, C., Pierce, S. A., Robson, B., Seppelt, R., Voinov, A. A., Fath, B. D., & Andreassian, V. (2013). ANDREASSIAN, V. Characterising performance of environmental models. *Environmental Modelling & Software*, *40*, 1-20. http://dx.doi.org/10.1016/j.envsoft.2012.09.011.

Beven, J. K. (2012). *Rainfall–Runoff Modelling: the primer*. (2nd ed., 488 p.). Chichester: John Wiley & Sons Ltd. http://dx.doi.org/10.1002/9781119951001.

Boyle, D. P., Gupta, H. V., & Sorooshian, S. (2000). Toward improved calibration of hydrologic models: combining the strengths of manual and automatic methods. *Water Resources Research*, *36*(12), 3663-3674. http://dx.doi.org/10.1029/2000WR900207.

Bravo, J. M., Paz, A. R., Collischonn, W., Uvo, C. B., Pedrollo, O. C., & Chou, S. C. (2009). Incorporating forecasts of rainfall in two hydrologic models used for medium-range streamflow forecasting. *Journal of Hydrologic Engineering*, *14*(5), 435-445. http://dx.doi.org/10.1061/(ASCE)HE.1943-5584.0000014.

Collischonn, W., Allasia, D., Silva, B. C., & Tucci, C. E. M. (2007). The MGB-IPH model for large-scale rainfall-runoff modelling. *Hydrological Sciences Journal*, *52*(5), 878-895. http://dx.doi.org/10.1623/hysj.52.5.878.

Criss, R. E., & Winston, W. E. (2008). Do Nash values have value? Discussion and alternate proposals. *Hydrological Processes*, *22*(14), 2723-2725. http://dx.doi.org/10.1002/hyp.7072.

Dakhlaoui, H., Bargaoui, A. Z., & Bárdossy, A. (2012). Toward a more efficient Calibration Schema for HBV rainfall-runoff model. *Journal of Hydrology (Amsterdam)*, *444-445*, 161-179. http://dx.doi.org/10.1016/j.jhydrol.2012.04.015.

Efstratiadis, A., & Koutsoyiannis, D. (2010). One decade of multi-objective calibration approaches in hydrological modelling: a review Multi-objective calibration approaches in hydrological modelling. *Hydrological Sciences Journal*, *55*(1), 58-78. http://dx.doi.org/10.1080/02626660903526292.

Felix, W. S., & Paz, A. R. (2016). Representação dos processos hidrológicos em bacia hidrográfica do semiárido paraibano com modelagem hidrológica distribuída. *Revista Brasileira de Recursos Hídricos- RBRH*, *21*(3), 556-569. https://doi.org/10.1590/2318-0331.011616009.

Fenícia, F., Solomatine, D. P., Savenije, H. H. G., & Matgen, P. (2007). Soft combination of local models in a multi-objective framework. *Hydrology and Earth System Sciences*, *11*(6), 1797-1809. http://dx.doi.org/10.5194/hess-11-1797-2007.

Fowler, K., Peel, M., Western, A., & Zhang, L. (2018). Improved rainfall-runoff calibration for drying climate: choice of objective function. *Water Resources Research*, *54*(5), 3392-3408. http://dx.doi.org/10.1029/2017WR022466.

Garcia, F., Folton, F., & Oudin, L. (2017). Which objective function to calibrate rainfall–runoff models for low-flow index simulations? *Hydrological Sciences Journal*, *62*(7), 1149-1166. http://dx.doi.org/10.1080/02626667.2017.1308511.

Gupta, H. V., Sorooshian, S., & Yapo, P. O. (1998). Toward improved calibration of hydrologic models: multiple and noncommensurable measures of information. *Water Resources Research*, *34*(4), 751-764. http://dx.doi.org/10.1029/97WR03495.

Gupta, H. V., Sorooshian, S., & Yapo, P. O. (1999). Status of automatic calibration for hydrologic models: comparison with multilevel expert calibration. *Journal of Hydrologic Engineering*, *4*(2), 135-143. http://dx.doi.org/10.1061/(ASCE)1084-0699(1999)4:2(135).

Gupta, H. V., Kling, H., Yilmaz, K., & Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: implications for improving hydrological modelling. *Journal of Hydrology (Amsterdam)*, *377*(1-2), 80-91. http://dx.doi.org/10.1016/j.jhydrol.2009.08.003.

Gutierrez, J. C. T., Adamatti, D. S., & Bravo, J. M. (2019). A new stopping criterion for multi-objective evolutionary algorithms: application in the calibration of a hydrologic model. *Computational Geosciences*, *23*(6), 1219-1235. http://dx.doi.org/10.1007/s10596-019-09870-3.

Herman, M. R., Nejadhashemi, A. P., Abouali, M., Hernandez-Suarez, J. S., Daneshvar, F., Zhang, Z., Anderson, M. C., Sadeghi, A. M., Hain, C. R., & Sharifi, A. (2018). Evaluating the role of evapotranspiration remote sensing data in improving hydrological modeling predictability. *Journal of Hydrology (Amsterdam)*, *556*, 39-49. http://dx.doi.org/10.1016/j.jhydrol.2017.11.009.

Hogue, T. S., Sorooshian, S., Gupta, H., Holz, A., & Braatz, D. (2000). A multistep automatic calibration scheme for river forecasting models. *Journal of Hydrometeorology*, *1*(6), 524-542. http://dx.doi.org/10.1175/1525-7541(2000)001<0524:AMACSF>2.0.CO;2.

Janssen, P. H. M., & Heuberger, P. S. C. (1995). Calibration of process-oriented models. *Ecological Modelling*, *83*(1-2), 55-66. http://dx.doi.org/10.1016/0304-3800(95)00084-9.

Kollat, J. B., Reed, P. M., & Wagener, T. (2012). When are multiobjective calibration trade-offs in hydrologic models meaningful? *Water Resources Research*, *48*(3), 1-19. http://dx.doi.org/10.1029/2011WR011534.

Krause, P., Boyle, D. P., & Base, F. (2005). Comparison of different efficiency criteria for hydrological model assessment. *Advances in Geosciences*, *5*, 89-97. http://dx.doi.org/10.5194/adgeo-5-89-2005.

Legates, D. R., & McCabe Junior, G. J. (1999). Evaluating the use of "goodness-of-fit" measures in hydrologicv and hydroclimatic model validation. *Water Resources Research*, *35*(1), 233-241. http://dx.doi.org/10.1029/1998WR900018.

Li, C., Wang, H., Liu, J., Yan, D., Yu, F., & Zhang, L. (2010). Effect of calibration data series length on performance and optimal parameters of hydrological model. *Water Science and Engineering*, *3*(4), 378-393.

Lindström, G., Johansson, B., Persson, M., Gardelin, M., & Bergström, S. (1997). Development and test of the distributed HBV-96 hydrological model. *Journal of Hydrology (Amsterdam)*, *201*(1-4), 272-288. http://dx.doi.org/10.1016/S0022-1694(97)00041-3.

Madsen, H. (2000). Automatic calibration of a conceptual rainfall–runoff model using multiple objectives. *Journal of Hydrology (Amsterdam)*, *235*(3-4), 276-288. http://dx.doi.org/10.1016/S0022-1694(00)00279-1.

Molina-Navarro, E., Andersen, H. E., Nielsen, A., Thodsen, H., & Trolle, D. (2017). The impact of the objective function in multi-site and multi-variable calibration of the SWAT model. *Environmental Modelling & Software*, *93*, 255-267. http://dx.doi.org/10.1016/j.envsoft.2017.03.018.

Moriasi, D. N., Arnold, J. G., Liew, M. W. V., Bingner, R. L., Harmel, R. D., & Veith, T. L. (2007). Model evaluation guidelines for systematic quatification of accuracy in watershed simulations. *American Society of Agricultural and Biological Engineers*, *50*(3), 885-900. http://dx.doi.org/10.13031/2013.23153.

Muleta, M. K. (2012). Model performance sensitivity to objective function during automated calibrations. *Journal of Hydrologic Engineering*, *17*(6), 756-767. http://dx.doi.org/10.1061/(ASCE)HE.1943-5584.0000497.

Nelson, A. M., Moriasi, D. N., Talebizadeh, M., Steiner, J. L., Confesor, R. B., Gowda, P. H., Starks, P. J., & Tadesse, H. (2017). Impact of length of dataset on streamflow calibration parameters and performance of apex model. *Journal of the American Water Resources Association*, *53*(5), 1164-1177. http://dx.doi.org/10.1111/1752-1688.12564.

Oliveira, R. F., Zolin, C. A., Victoria, D. C., Lopes, T. R., Vendrusculo, L. G., & Paulino, J. (2018). Hydrological calibration and validation of the MGBIPH model for water resource management in the upper Teles Pires River basin in the Amazon-Cerrado ecotone in Brazil. *Acta Amazonica*, *49*(1), 54-63. http://dx.doi.org/10.1590/1809-4392201800812.

Paiva, R. C. D., Buarque, D. C., Collischonn, W., Bonnet, M.-P., Frappart, F., Calmant, S., & Mendes, C. A. B. (2013). Largescale hydrologic and hydrodynamic modeling of the Amazon River basin. *Water Resources Research*, *49*(3), 1226-1243. http://dx.doi.org/10.1002/wrcr.20067.

Pappenberger, F., & Beven, K. J. (2004). Functional classification and evaluation of hydrographs based on Multicomponent Mapping (Mx). *Intl. J. River Basin Management*, *2*(2), 89-100. http://dx.doi.org/10.1080/15715124.2004.9635224.

Parker, S. R., Adams, S. K., Lammers, R. W., Stein, E. D., & Bledsoe, B. P. (2019). Targeted hydrologic model calibration to improve prediction of ecologically-relevant flow metrics. *Journal of Hydrology (Amsterdam)*, *573*, 546-556. http://dx.doi.org/10.1016/j.jhydrol.2019.03.081.

Pechlivanidis, I. G., Jackson, B. M., Mcmillan, H. K., & Gupta, H. V. (2012). Using an informational entropy-based metric as a diagnostic of flow duration to drive model parameter identification. *Journal Global NEST*, *14*(3), 325-334.

Pereira, F. F., Moraes, M. A. E., & Uvo, C. B. (2014). Implementation of a two-way coupled atmosphericydrological system for environmental modeling at regional scale. *Hydrology Research*, *45*(3), 504-514. http://dx.doi.org/10.2166/nh.2013.335.

Pool, S., Vis, M., & Seibert, J. (2018). Evaluating model performance:towards a non-parametric variant of the Kling-Gupta efficiency. *Hydrological Sciences Journal*, 63:13-14, 1941-1953. https://doi.org/10.1080/02626667.2018.1552002.

Pushpalatha, R., Perrin, C., Moine, N. L., & Andréassian, V. (2012). A review of efficiency criteria suitable for evaluating low-flow simulations. *Journal of Hydrology (Amsterdam)*, *420-421*, 171-182. http://dx.doi.org/10.1016/j.jhydrol.2011.11.055.

Rajib, M. A., Merwade, V., & Yu, Z. (2016). YU; Z. Multi-objective calibration of a hydrologic model using spatially distributed emotely sensed/in-situ soil moisture. *Journal of Hydrology (Amsterdam)*, *536*, 192-207. http://dx.doi.org/10.1016/j.jhydrol.2016.02.037.

Ribeiro Neto, A., Collischonn, W., Silva, R. C. V., & Tucci, C. E. M. (2006). Hydrological modelling in Amazonia—use of the MGB-IPH model and alternative databases. In M. Sivapalan, T. Wagener, S. Uhlenbrook, E. Zehe, V. Lakshmi, X. Liang, Y. Tachikawa & P. Kumar. *Predictions in ungauged basins*: promise and progress (IAHS Series of Proceedings and Reports, Vol. 303, pp. 246-254). Wallingford: IAHS Publications.

Rientjes, T. H. M., Muthuwatta, L. P., Bos, M. G., Booij, M. J., & Bhatti, H. A. (2013). Multi-variable calibration of a semi-distributed hydrological model using streamflow data and satellite-based evapotranspiration. *Journal of Hydrology (Amsterdam)*, *505*, 276-290. http://dx.doi.org/10.1016/j.jhydrol.2013.10.006.

Romanowicz, R. J., Osuch, M., & Grabowiecka, M. (2013). On the choice of calibration periods and objective functions: a practical guide to model parameter identification. *Acta Geophysica*, *61*(6), 1477-1503. http://dx.doi.org/10.2478/s11600-013-0157-6.

Rwetabula, J., De Smedt, F., & Rebhun, M. (2012). Simulation of hydrological processes in the Simiyu River, tributary of Lake Victoria, Tanzania. *Water AS*, *38*(4), 623-632. http://dx.doi.org/10.4314/wsa.v38i4.18.

Schaefli, B., & Gupta, H. V. (2007). Do Nash values have value? *Hydrological Processes*, *21*(15), 2075-2080. http://dx.doi.org/10.1002/hyp.6825.

Troin, M., Arsenault, R., & Brissette, F. (2015). Performance and Uncertainty Evaluation of Snow Models on Snowmelt Flow Simulations over a Nordic Catchment (Mistassibi, Canada). *Hydrology*, *2*(4), 289-317. http://dx.doi.org/10.3390/hydrology2040289.

Tucci, C. E. M. (2005). *Modelos hidrológicos* (2. ed., 678 p.). Porto Alegre: Editora da UFRGS ABRH GWP.

Tucci, C. E. M., Collischonn, W., Clarke, R. T., Paz, A. R., & Allasia, D. (2008). Short- and long-term flow forecasting in the Rio Grande watershed (Brazil). *Atmospheric Science Letters*, *9*(2), 53-56. http://dx.doi.org/10.1002/asl.165.

Tucci, C. E. M., Marengo, J. A., Dias, P. S., Collischonn, W., Silva, B., Clarke, R., Cardoso, A., Negrón-Juárez, R., Sampaio, G., & Chou, S. C. (2005). *Streamflow forecasting in São Francisco River basin based on climatic forecasting* (Technical Report ANEEL/WMO/98/00). Porto Alegre: Editora.

Van Liew, M. W., Veith, T. L., Bosch, D. D., & Arnold, J. G. (2007). Suitability of SWAT for the Conservation Effects Assessment Project: Comparison on USDA Agricultural Research Service Watersheds. *Journal of Hydrological Research*, *12*(2), 173-189. http://dx.doi.org/10.1061/(ASCE)1084-0699(2007)12:2(173).

Westerberg, I. K., Guerrero, J. L., Younger, P. M., Beven, K. J., Seibert, J., Halldin, S., Freer, J. E., & Xu, C.-Y. (2011). Calibration of hydrological models using flow-duration curves. *Hydrology and Earth System Sciences*, *15*(7), 2205-2227. http://dx.doi.org/10.5194/hess-15-2205-2011.

Wohling, T., Samaniego, L., & Kumar, R. (2013). Evaluating multiple performance criteria to calibrate the distributed hydrological model of the upper Neckar catchment. *Environmental Earth Sciences*, *69*(2), 453-468. http://dx.doi.org/10.1007/s12665-013-2306-2.

Zhang, Y., Shao, Q., Zhang, S., Zhai, X., & She, D. (2016). Multi-metric calibration of hydrological model to capture overall flow regimes. *Journal of Hydrology (Amsterdam)*, *539*, 525-538. http://dx.doi.org/10.1016/j.jhydrol.2016.05.053.

Zink, M., Mai, J., Cuntz, M., & Samaniego, L. (2018). Conditioning a hydrologic model using patterns of remotely sensed land surface temperature. *Water Resources Research*, *54*(4), 2976-2998. http://dx.doi.org/10.1002/2017WR021346.

Authors contributions

Paloma Mara de Lima Ferreira: Literature review, study design, methodology development and application, results discussion and paper writing.

Adriano Rolim da Paz: Study design, results discussion and paper writing.

Juan Martín Bravo: Results discussion and paper writing.