Title: Decomposition of the Mean Squared Error & NSE Performance Criteria: Implications for Improving Hydrological Modelling

Keywords: mean squared error; Nash-Sutcliffe efficiency; model performance evaluation; calibration; multiple criteria; hydrologic modelling; criteria decomposition; diagnostic analysis

Corresponding Author: Dr. Harald Kling, Ph.D.
Corresponding Author's Institution: The University of Arizona
First Author: Hoshin V Gupta
Order of Authors: Hoshin V Gupta; Harald Kling, Ph.D.; Koray K Yilmaz; Guillermo F Martinez-Baquero
Suggested Reviewers: Bettina Schaefli PhD
Assistant Prof., Water Resources, TU Delft
b.schaefli@tudelft.nl
related paper on NSE

Doug Boyle PhD
Assistant Prof., Desert Research Institute
Doug.Boyle@dri.edu
experience in multi-criteria analysis

Vazken Andreassian PhD
Head of Research Unit, Cemagref
vazken.andreassian@cemagref.fr
extensive experience in calibration with NSE

Peter Krause PhD
Assistant Prof., Geoinformatik Jena, Friedrich-Schiller-Universität Jena
p.krause@uni-jena.de
related paper comparing different performance criteria

Andras Bardossy PhD
Prof., Institut für Wasserbau, Universität Stuttgart
Andras.Bardossy@iws.uni-stuttgart.de
experience in regional modelling and evaluation
April 7, 2009

Konstantine Georgakakos, Editor
Journal of Hydrology
Hydrologic Research Center
San Diego, CA, USA

Dear Mr. Georgakakos,


Best Regards
Harald Kling, PhD
Department of Hydrology and Water Resources
The University of Arizona,
Tucson, AZ 85721, USA
Phone: 520-626-9712
Email: harald.kling@boku.ac.at
Title:

Decomposition of the Mean Squared Error & NSE Performance Criteria: Implications for Improving Hydrological Modelling

Authors:

Hoshin V. Gupta¹, Harald Kling¹, Koray K. Yilmaz², Guillermo F. Martinez-Baquero¹

Affiliations:

1: Department of Hydrology and Water Resources, The University of Arizona, 1133 E North Campus Dr, Tucson, AZ 85721, USA

2: now at Earth System Science Interdisciplinary Center, University of Maryland, College Park, MD 20742, USA and at NASA Goddard Space Flight Center, Laboratory for Atmospheres, Greenbelt, MD 20771, USA

Corresponding author:

Harald Kling
e-mail: harald.kling@boku.ac.at
Phone: +1 520 626 9712
Fax: +1 520 621 1422
Abstract:

The mean squared error (MSE) and the related normalization, the Nash-Sutcliffe efficiency (NSE), are the two criteria most widely used for calibration and evaluation of hydrological models with observed data. Here, we present a diagnostically interesting decomposition of NSE (and hence MSE), which facilitates analysis of the relative importance of its different components in the context of hydrological modelling, and show how model calibration problems can arise due to interactions among these components. The analysis is illustrated by calibrating a simple conceptual precipitation-runoff model to daily data for a number of Austrian basins having a broad range of hydro-meteorological characteristics. Evaluation of the results clearly demonstrates the problems that can be associated with any calibration based on the NSE (or MSE) criterion. While we propose and test an alternative criterion that can help to reduce model calibration problems, the primary purpose of this study is not to present an improved measure of model performance. Instead, we seek to show that there are systematic problems inherent with any optimization based on formulations related to the MSE. The analysis and results have implications to the manner in which we calibrate and evaluate environmental models; we discuss these and suggest possible ways forward that may move us towards an improved and diagnostically meaningful approach to model performance evaluation and identification.

Keywords:

mean squared error; Nash-Sutcliffe efficiency; model performance evaluation; calibration; multiple criteria; hydrologic modelling; criteria decomposition; diagnostic analysis
1 Introduction

The mean squared error (MSE) criterion and its related normalization, the Nash-Sutcliffe efficiency (NSE, defined by Nash and Sutcliffe 1970) are the two criteria most widely used for calibration and evaluation of hydrological models with observed data. The value of MSE depends on the units of the predicted variable and varies on the interval [0.0 to inf], whereas NSE is dimensionless, being scaled onto the interval [-inf to 1.0]. As a consequence, the NSE value - obtained by dividing MSE by the variance of the observations and subtracting that ratio from 1.0 (Eq. 1 and Eq. 2) - is commonly the measure of choice for reporting (and comparing) model performance. Further, NSE can be interpreted as a classic skill score (Murphy 1988), where ‘skill’ is interpreted as the comparative ability of a model with regards to a baseline ‘model’, which in the case of NSE is taken to be the ‘mean of the observations’ (i.e., if NSE ≤ 0, the model is no better than using the observed mean as a predictor). The equations are:

\[
MSE = \frac{1}{n} \sum_{t=1}^{n} (x_{s,t} - x_{o,t})^2 \quad \text{Eq. 1}
\]

\[
NSE = 1 - \frac{\sum_{t=1}^{n} (x_{s,t} - x_{o,t})^2}{\sum_{t=1}^{n} (x_{o,t} - \mu_o)^2} = 1 - \frac{MSE}{\sigma_o^2} \quad \text{Eq. 2}
\]

where \( n \) is the total number of time-steps, \( x_{s,t} \) is the simulated value at time-step \( t \), \( x_{o,t} \) is the observed value at time-step \( t \), and \( \mu_o \) and \( \sigma_o \) are the mean and standard deviation of the observed values. In optimization MSE is subject to minimization and NSE is subject to maximization.
While the NSE criterion may be a convenient and popular (albeit gross) indicator of model skill, there has been a long and vivid discussion about the suitability of NSE (McCuen and Snyder 1975, Martinec and Rango 1989, Legates and McCabe 1999, Krause et al. 2005, McCuen et al. 2006, Schaefli and Gupta 2007) and several authors have proposed modifications - e.g. Mathevet et al. (2006) proposed a bounded version of NSE and Criss and Winston (2008) proposed a volumetric efficiency to be used instead of NSE. One of the main concerns about NSE is its use of the observed mean as baseline, which can lead to overestimation of model skill for highly seasonal variables such as runoff in snowmelt dominated basins. A comparison of NSE across basins with different seasonality (as is often reported in the literature) should therefore be interpreted with caution. For such situations, various authors have recommended the use of the seasonal or climatological mean as a baseline model (Garrick et al. 1978, Murphy 1988, Martinec and Rango 1989, Legates and McCabe 1999, Schaefli and Gupta 2007).

It is now generally accepted that the calibration of hydrological models should be approached as a multi-objective problem (Gupta et al. 1998). Within a multiple-criteria framework, the MSE and NSE criteria continue to be commonly used, because they can be computed separately for (1) different types of observations (e.g. runoff and snow observations; Bergström et al. 2002), (2) different locations (e.g. runoff at multiple gauges; Madsen 2003), or (3) different subsets of the same observation (e.g. rising and falling limb of the hydrograph; Boyle et al. 2000). More generally, however, different types of model performance criteria - such as NSE, coefficient of correlation, bias, etc. - can be computed from multiple variables and/or at multiple sites (see Anderton et al. 2002, Beldring 2002, Rojanschi et al. 2005, Cao et al. 2006, and others).

When handled in this manner, the model calibration problem can be treated as a full multiple-criteria optimization problem resulting in a ‘Pareto set’ of non-dominated solutions (Gupta et al 1998), or reduced to a related single-criterion optimization problem by combining the different
(weighted) criteria into one overall objective function. Numerous examples of the latter approach exist in the literature where NSE or MSE appear in an overall objective function (e.g. Lindström 1997, Bergström et al. 2002, Madsen 2003, van Griensven and Bauwens 2003, Parajka et al. 2005, Young 2006, Rode et al. 2007, Marce et al. 2008, Wang et al. 2009), because it conveniently enables the application of efficient single-criterion automated search algorithms, such as SCE (Shuffled Complex Evolution, Duan et al. 1992) or DDS (Dynamically Dimensioned Search, Tolson and Shoemaker 2007).

When using multiple criteria in evaluation, it has to be considered that some of these criteria are mathematically related, which is not always recognized (Weglarczyk 1998). For example, it is possible to decompose the NSE criterion into separate components, as shown by Murphy (1988) and Weglarczyk (1998), which facilitates a better understanding of how different criteria are interrelated and thereby enable more insight into what is causing a particular model performance to be ‘good’ or ‘bad’. Equally important, the decomposition can provide insight into possible trade-offs between the different components.

In this paper we present a diagnostically interesting decomposition of NSE (and hence MSE), which facilitates analysis of the relative importance of different components in the context of hydrological modelling, and show how model calibration problems can arise due to interactions among these components. Based on this analysis, we propose and test alternative criteria that can help to avoid these problems. The analysis is illustrated by calibrating a simple precipitation-runoff model to daily data for a number of Austrian basins having a broad range of hydro-meteorological characteristics, and evaluating the results on both the calibration and an independent ‘evaluation’ period. The results clearly demonstrate the problems that can be associated with any calibration based on the NSE (or MSE) criterion. The analysis and results have interesting implications to the
manner in which we calibrate and evaluate environmental models; we discuss these and some possible ways forward in the discussion and conclusions sections.

2 Decomposition of model performance criteria

2.1 Decomposition of NSE

A decomposition of criteria based on mean squared errors reveals that there are three distinctive components, represented by the correlation, the conditional bias, and the unconditional bias, as evident in Eq. 3 which shows a decomposition of NSE (Murphy 1988, Weglarczyk 1998).

\[ \text{NSE} = A - B - C \quad \text{Eq. 3} \]

with:

\[ A = r^2 \]

\[ B = \left[ r - (\sigma_s / \sigma_o) \right]^2 \]

\[ C = \left[ (\mu_s - \mu_o) / \sigma_o \right]^2 \]

where \( r \) is the linear correlation coefficient between \( x_s \) and \( x_o \), and \((\mu_s, \sigma_s) \) and \((\mu_o, \sigma_o) \) represent the first two statistical moments (means and standard deviations) of \( x_s \) and \( x_o \) respectively. The quantity \( A \) measures the strength of the linear relationship between the simulated and observed values, \( B \) measures the conditional bias, and \( C \) measures the unconditional bias (Murphy 1988).

However, an alternative way in which to reformulate Eq. 3 is given below as Eq. 4.

\[ \text{NSE} = 2 \cdot \alpha \cdot r - \alpha^2 - \beta_o^2 \quad \text{Eq. 4} \]

with:

\[ \alpha = \sigma_s / \sigma_o \]
where the quantity $\alpha$ is a measure of relative variability in the simulated and observed values, and $\beta_n$ is the bias normalized by the standard deviation in the observed values (note that $\beta_n = \sqrt{C}$).

Eq. 4 shows that NSE is composed of three components, two of which relate to the ability of the model to reproduce the first and second moments of the distribution of the observations (i.e. mean and standard deviation), while the third relates to the ability of the model to reproduce timing and shape as measured by the correlation coefficient. The ideal values for the three components are $r = 1$, $\alpha = 1$, and $\beta_n = 0$. From a hydrological perspective, 'good' values for each of these three components are highly desirable, since in general we aim at matching the overall volume of flow, the spread of flows (e.g. flow duration curve), and the timing and shape of (for example) the hydrograph (Yilmaz et al. 2008). It is clear, therefore, that optimizing NSE is essentially a search for a balanced solution among the three components, which is similar to the multiple-criteria approach of computing an overall (weighted) objective function from several different criteria as discussed in the introduction.

However, in using NSE we must be concerned with two facts. First, the bias $(\mu_s - \mu_o)$ component appears in a normalized form, scaled by the standard deviation in the observed flows. This means that in basins with high runoff variability the bias component will tend to have a smaller contribution (and therefore impact) in the computation and optimization of NSE, possibly leading to model simulations having large volume balance errors. In a multiple-criteria sense, this is equivalent to using a weighted objective function with a low weight applied to the bias component.

Second, and equally serious, the quantity $\alpha$ appears twice in Eq. 4, exhibiting an interesting (and problematic) interplay with the linear correlation coefficient $r$. It is easy to show, by taking the first derivative of NSE (in Eq. 4) with respect to $\alpha$, that the maximum value of NSE is obtained when
\[ \alpha = r. \] And, since \( r \) will always be smaller than unity, this means that in maximizing NSE we will tend to select a value for \( \alpha \) that underestimates the variability in the flows (more precisely, we will favour models/parameter sets that generate simulated flows that underestimate the variability).

Taking these two facts together, we note that when \( \beta_s = 0 \) and \( \alpha = r \), then the NSE is equivalent to \( r^2 \), which is the well-known coefficient of determination. Therefore, \( r^2 \) can be interpreted as a maximum (potential) value for NSE if the other two components are able to achieve their ‘optimal’ values.

Fig. 1 illustrates the relationship of NSE with \( r \) and \( \alpha \), while assuming that \( \beta_s \) is zero (\( \beta_s \) is only an additive term, anyway). For a given \( r \) the ‘optimal’ \( \alpha \) for maximizing NSE lies on the 1:1 line, although the ideal value of \( \alpha \) is on a horizontal line at 1.0. This theoretical relationship is illustrated in Fig. 1a. Of course, not all combinations of \( r \) and \( \alpha \) may be possible with a hydrological model due to restrictions imposed by the model structure, feasible parameter values and input-output data.

However, Fig. 1b shows a real example in which random sampling of the parameter space actually seems to cover a large portion of the theoretical criteria space. Since the model used here (HyMod model, Boyle 2000) is a simple, but representative, example of watershed models in common use, the problematic interplay between \( \alpha \) and \( r \) is likely to be of importance for any type of hydrological model that is optimized with NSE.

Further, the same exact problems will arise when using MSE as a model calibration criterion. We can substitute Eq. 4 into Eq. 2, and thereby obtain Eq. 5 which shows the related decomposition of the MSE criterion, consisting (again) of three error terms, but here all three of them are additive.

\[
MSE = 2 \cdot \sigma_s \cdot \sigma_o \cdot (1 - r) + (\sigma_s - \sigma_o)^2 + (\mu_s - \mu_o)^2 \quad \text{Eq. 5}
\]
From Eq. 3, Eq. 4 and Eq. 5 it should be immediately obvious that many different combinations of
the three components can result in the same overall value for MSE or NSE, respectively, potentially
leading to considerable ambiguity in the comparative evaluation of alternative model hypotheses.
The relative contribution of each of these components to the overall MSE can be computed as:

\[ f_i = \frac{F_i}{\sum_{j=1}^{3} F_j} \quad \text{Eq. 6} \]

with:

\[ F_1 = 2 \cdot \sigma_s \cdot \sigma_o \cdot (1 - r) \]
\[ F_2 = (\sigma_s - \sigma_o)^2 \]
\[ F_3 = (\mu_s - \mu_o)^2 \]

### 2.2 Alternative model performance criteria

As discussed above, a peculiar feature of the NSE criterion is the problematic interplay between \( \alpha \)
and \( r \), which is likely to result in an underestimation of the variability in the flows. One way to
overcome this is by inflating the observed variability as indicated by Eq. 7, while at the same time
preserving the mean of the observations and their linear correlation with the simulations. Using Eq.
7 with Eq. 4 results in Eq. 8, which represents a ‘corrected’ version of NSE:

\[ x_{o,t}^* = x_{o,t} - \mu_o + \mu_o \quad \text{Eq. 7} \]

\[ \text{NSE}_{corr} = \frac{1}{c} \cdot 2 \cdot \alpha \cdot r - \frac{1}{\alpha^2} \cdot \alpha^2 - \frac{1}{\alpha^2} \cdot \beta_s^2 \quad \text{Eq. 8} \]
where $c$ is correction factor to inflate the variability in the observed flows. It can be easily shown
that if $c$ is set equal to $1/r$, it will assure that a value of $\alpha = 1$ will now maximize $\text{NSE}_{\text{cor}}$ (as opposed to $\alpha = r$ maximizing $\text{NSE}$).

Alternatively, instead of trying to come up with a ‘corrected’ $\text{NSE}$ criterion, since MSE and $\text{NSE}$
can be decomposed into three components, the whole calibration problem can instead be viewed
from the multi-objective perspective, by focusing on the correlation, variability error and bias error
as separate criteria to be optimized. In doing this, it makes sense to enable a better hydrological
interpretation of the bias component by using the ratio of the means of the simulated and observed
flows ($\beta$) for this further analysis - as opposed to using $\beta_0$. With this formulation, using $\beta$ instead of
$\beta_0$, all three of the components now have their ideal value at unity.

Fig. 2 shows an example for the trade-off between the three components for a simple hydrological
model using random parameter sampling. The plot shows a distinctive Pareto front in the three-
dimensional criteria space. If it is desired to select a compromise solution from the Pareto front, one
possible approach is to compute for all points the Euclidian distance from the ideal point and then to
subsequently select the point having the shortest distance (Eq. 9). Since all three of the components
are dimensionless numbers, we are able to obtain a reasonable solution for the Euclidian distance in
the un-transformed criteria space. Alternatively, a re-scaling of the axes in the criteria space is
easily obtained via Eq. 10. In this paper, we will only explore the use of the KGE criterion (Eq. 9),
which is equivalent to setting all three scaling factors of Eq. 10 to unity.

**Fig. 2 near here**

\[
KGE = 1 - \frac{ED}{\text{Eq. 9}}
\]

\[
KGE_s = 1 - \frac{ED_s}{\text{Eq. 10}}
\]
with: \[ ED = \sqrt{(r-1)^2 + (\alpha - 1)^2 + (\beta - 1)^2} \]

\[ ED_s = \sqrt{[s_r \cdot (r-1)]^2 + [s_{\alpha} \cdot (\alpha - 1)]^2 + [s_{\beta} \cdot (\beta - 1)]^2} \]

\[ \beta = \frac{\mu_s}{\mu_e} \]

where \( ED \) is the Euclidean distance from the ideal point, \( ED_s \) is the Euclidean distance from the ideal point in the scaled space, \( \beta \) is the ratio between the mean simulated and mean observed flows, i.e. \( \beta \) represents the bias; \( s_r, s_{\alpha} \) and \( s_{\beta} \) are scaling factors that can be used to re-scale the criteria space before computing the Euclidean distance from the ideal point, i.e. \( s_r, s_{\alpha} \) and \( s_{\beta} \) can be used for adjusting the emphasis on different components.

Analogous to Eq. 6 we can compute the relative contribution of the three components with Eq. 11.

\[ g_j = G_j / \sum_{j=1}^{3} G_j \quad \text{Eq. 11} \]

with: \[ G_1 = (r-1)^2 \]

\[ G_2 = (\alpha - 1)^2 \]

\[ G_3 = (\beta - 1)^2 \]

2.3 Notes on regression lines

As is well known, the slope of the regression lines and the coefficient of correlation are related (Eq. 12 to Eq. 14). Since different ‘optimal’ values for \( \alpha \) are obtained by the NSE and KGE criteria, this also leads to implications for the regression lines.
\[ r^2 = k_s \cdot k_o \quad \text{Eq. 12} \]

\[ k_s = \frac{\text{Cov}_{so}}{\sigma_s^2} = \frac{r}{\alpha} \quad \text{Eq. 13} \]

\[ k_o = \frac{\text{Cov}_{so}}{\sigma_o^2} = r \cdot \alpha \quad \text{Eq. 14} \]

with: \[ r = \frac{\text{Cov}_{so}}{\sigma_s \cdot \sigma_o} \]

where \( \text{Cov}_{so} \) is the covariance between the simulated and observed values, \( k_s \) is the slope of the regression line when regressing the observed against the simulated values, and \( k_o \) is the slope of the regression line when regressing the simulated against the observed values.

Murphy (1988) has already noted that for NSE the conditional bias term \( B \) in Eq. 3 will vanish only if the slope of the regression line \( k_s \) is equal to unity (i.e. regressing the observed against the simulated values), which is desirable in the context of the ‘verification’ of forecasts. This means that for a given forecast (simulated value), the expected value of the observed value lies on the 1:1 line (assuming a Gaussian distribution). As discussed before, the optimal value of \( \alpha \) that maximizes NSE is given by a model simulation for which \( \alpha \) is equal to \( r \). As evident in Eq. 13 this results in \( k_s = 1 \), but at the same time this also implies that \( k_o = r^2 \) (Eq. 14). Because \( r^2 \) will always be smaller than unity, this means that we will, in general, tend to underestimate the slope of the regression line when regressing the simulated against the observed values. The tendency will be for high values (peak flows) to be underestimated and for low values (recessions) to be overestimated in the simulation.
In brief, for maximizing NSE the optimal values for \( k \) and \( k_o \) are unity and \( r^2 \), respectively. In the case of KGE, the optimal value for \( \alpha \) is at unity, which means that for maximizing KGE the optimal values for both \( k \) and \( k_o \) are equal to \( r \). Again, since \( r \) is smaller than unity we will tend to underestimate high values and overestimate low values.

In considering this, it should be noted that both approaches for computing the regression lines (regressing observed against simulated values, or vice versa) are valid, but have different interpretations. In the context of runoff simulations, when using \( k \), we are basing the evaluation on the expected error in simulation of the observed runoff being zero for a given simulated runoff, which is a sensible approach when making runoff forecasts under ‘normal’ conditions. However, if we are interested in the ‘unusual’ runoff conditions - such as runoff peaks - then a more sensible approach would be to use \( k_o \), where we are interested in the question, "If a flood occurs, can we forecast (simulate) it?", whereas in the case of \( k \) such a runoff peak is ‘averaged out’. Fig. 3 illustrates this with typical scatter plots for runoff simulation. In this example, \( k \) is close to unity, suggesting unbiased forecasts (Fig. 3a), and at the highest simulated flows of around 10 m\(^3\)/s the small number of observed flows (runoff peaks) that are well above the regression line are ‘averaged out’ by the larger number of observed flows that occur slightly below the regression line. However, it is clear that whenever a runoff peak above 10 m\(^3\)/s occurs, there is a clear tendency for underestimation in the simulation (Fig. 3b).

These problems arise because the distribution of runoff is usually highly skewed. If \( k_o \) is of higher interest, then the use of NSE may cause problems, since the simulated runoff will tend to underestimate the peak flows. In the case of the KGE criterion, we will also have a tendency towards underestimation, but not as severe as with the NSE. Note that for extreme low-flows, similar considerations as for the runoff peaks apply (but here we will tend to overestimate the low-flow).
Case Study

To examine and illustrate the implications of the theoretical considerations presented above we applied a simple conceptual precipitation-runoff model to several basins. Using NSE (Eq. 2) and KGE (Eq. 9) as model performance criteria, two different sets of parameters were obtained for each basin by calibration against observed runoff data. For each parameter set we compare the overall model performance as evaluated by the NSE and KGE criteria and, in addition, conduct a detailed analysis of the criterion components. Further, we also examine the model performance on an independent ‘evaluation’ period.

3.1 Study area

For this study we used the forty-nine mesoscale Austrian basins (Fig. 4) used in the regionalization study reported by Kling and Gupta (2009). All are pre-alpine or lowland basins where snowmelt does not dominate runoff generation. They vary in size from 112.9 km$^2$ to 689.4 km$^2$, with a median size of 287.3 km$^2$, and a mean elevation range from 232 m to 952 m above sea level. The basins represent a wide range of physiographic and meteorological properties, with the most important land-use types being forest, grassland and agriculture. According to the Hydrological Atlas of Austria (BMLFUW 2007), the long-term mean annual precipitation in the basins ranges from 507 to 1929 mm, and the corresponding runoff ranges from 44 to 1387 mm, resulting in a large range of runoff coefficients (from 9 to 72 percent). Thus, both wet and dry basins are included. Fig. 5 shows a diagnostic plot where normalized actual evapotranspiration is plotted against normalized precipitation (both variables are scaled by potential evapotranspiration); it indicates that most of the basins are energy limited and only a few of the basins are water limited.
3.2 Data basis

We used observed daily data for the period September 1990 to August 2000; the first two years were used as a warm-up period, the next five years for calibration, and the final three years for independent evaluation. Observed catchment outlet runoff data were used for parameter calibration in each of the basins. Precipitation inputs were based on daily data from 222 stations, regionalized using the method of Thiessen-Polygons. Air temperature inputs were based on data from 98 stations, regionalized via linear regression with elevation. Potential evapotranspiration inputs were based on monthly fields of potential evapotranspiration (Kling et al. 2007) with a spatial resolution of 1x1 km. The monthly potential evapotranspiration data were disaggregated to daily time-steps by using daily data from 21 indicator stations, where the daily potential evapotranspiration was computed using the Thornthwaite-method (Thornthwaite and Mather 1957).

3.3 Hydrological model

A simple, conceptual, spatially distributed daily precipitation-runoff model similar to the HBV model (Bergström 1995) was used; the model was previously applied to these same basins by Kling and Gupta (2009). The model uses a 1x1 km² raster grid for spatial discretization of the basins. However, for simplicity, the current study assumes uniform parameter fields. Inputs to the model are precipitation, air temperature, and potential evapotranspiration. The model consists of a snow module, soil moisture accounting, runoff separation into different components, and a routing module. Snowfall is determined from precipitation data using a threshold temperature, and snowmelt is computed with the temperature-index method (see e.g. Hock 2003). Rainfall and
snowmelt are input to the soil module, where runoff generation is computed via an exponential formulation that accounts for current soil moisture conditions (see e.g. Bergström and Graham 1998). Actual evapotranspiration depletes the soil moisture store; the rate of actual evapotranspiration depends on current soil moisture conditions and potential evapotranspiration. Runoff is separated into fast (surface flow) and slow (base flow) components by two linear reservoirs having different recession coefficients. A further linear reservoir is used to simulate channel routing of the runoff. Fig. 6 shows the conceptual structure of the model (the snow module is not shown). The model equations are presented in Kling and Gupta (2009). Table 1 lists the most important parameters of the model.

To reduce dimensionality of the parameter calibration problem, some of the model parameters are set to plausible values and are not further calibrated. This applies to snow parameters, because snow is of limited importance in the basins of this study, and to the channel routing parameters, which are of limited importance at the daily time-step (the values of Kling and Gupta (2009) are used). In addition, the critical soil moisture for reducing actual evapotranspiration is set to a constant value. The six remaining parameters were calibrated using the Shuffled Complex Evolution optimization algorithm (SCE, Duan et al. 1992), using six complexes.

Fig. 6 near here

Table 1 near here

### 3.4 Results

The optimization runs resulted in two parameter sets for each basin. Optimization using the ‘optNSE’ method results in parameter sets ‘$\theta_{\text{optNSE}}$’ that yield optimal runoff simulations maximizing NSE (Eq. 4), while optimization using the ‘optKGE’ method results in parameter sets ‘$\theta_{\text{optKGE}}$’ that yield optimal runoff simulations maximizing KGE (Eq. 9). A standard method for
reporting model performance in precipitation-runoff modelling studies is to present scatter plots of NSE between calibration and evaluation periods (see e.g. Merz and Blöschl 2004). Fig. 7 displays such a scatter plot; as expected, for many basins the NSE deteriorates when going from the calibration to the evaluation period (Fig. 7a). Similar results are obtained for KGE (Fig. 7b).

Now, there can be different reasons for deterioration of model performance on the evaluation period. These include over-fitting of the parameters to the calibration period, non-stationarity between the calibration and evaluation periods, lack of ‘power’ in the objective function, etc. Instead of falsification and model rejection, which would be a logical conclusion from such a result, it is common practice to simply report the deterioration in the model performance and then to move on. In our case, we can report that when moving from calibration to evaluation period the median NSE has decreased from 0.76 to 0.59 and the median KGE has decreased from 0.86 to 0.72, but what hydrological meaning do these numbers have? Here, an analysis of the different components that constitute the overall model performance enables us to learn much more about the model behaviour, differences between the calibration and evaluation periods, and also differences between basins.

Before analysing the criterion components it is interesting to note the relationship between NSE and KGE. Fig. 7 shows that when optimizing on KGE (optKGE) there is a strong correlation between the values obtained for the KGE and NSE criteria (Fig. 7d). However, when optimizing on NSE (optNSE), the correlation between the values obtained for NSE and KGE is lower (Fig. 7c). The reasons for this will become much clearer later in this section, but briefly it is useful to keep in mind that optimization on KGE strongly controls the values that the $\alpha$ and $\beta$ components can achieve, whereas optimization on NSE constrains these components only weakly.
The relative contributions of the criterion components to the overall model performance obtained via optimization are shown in Fig. 8 (see Eq. 6 for optNSE and Eq. 11 for optKGE). The obtained (optimized) model performance is dominated by the component representing \( r \) (dark grey), whereas the other components representing the bias (light grey) and the variability (medium grey) of flows have only small relative contributions. This applies for all 49 basins and for both optimization on NSE (Fig. 8a) and KGE (Fig. 8b). However, a low relative contribution of a component to the final value of the (optimized) model performance does not necessarily imply that the model performance criterion is, in general, insensitive to this component. Instead, the relative contribution of a component can be small because of (1) low ‘weight’ of the component in the equation for calculating the overall model performance, and/or (2) the value of the component is close to its optimal value. As a consequence of (2), the relative contribution of the components representing the bias and the variability of flows can become large for non-optimal parameter sets.

To illustrate these considerations, Fig. 8c and Fig. 8d show the relative contribution of the criterion components using random parameter sampling for a selected basin (Glan River). The sampled points are arranged from left to right in order of decreasing performance for the selected criterion. With decreasing overall model performance (either NSE or KGE) there is a general tendency for the relative contribution of \( r \) to decrease and for the other two components to become much more important. In some cases only the component representing the bias is dominant, whereas in other cases only the component representing the variability of flows is dominant. This clearly indicates that both NSE and KGE are sensitive to all three of the components. From a multi-objective point of view this is definitely desirable, because it means that by calibrating on the overall model performance we can substantially improve the components representing the bias and the variability of flows. Here of course we should remember that in NSE the bias is normalized by the standard deviation.
deviation of the observed flows and that the ‘optimal’ $a$ is equal to $r$. Hence, with NSE it is not
necessarily assured that from a hydrological point of view good values for $a$ and $\beta$ are obtained.

Fig. 8 near here

The cumulative distribution functions for the NSE, $r$, $\alpha$, and $\beta$ measures as obtained with optNSE
and optKGE in the calibration and evaluation periods are shown in Fig. 9. Looking first at the
results for the NSE criterion (Fig. 9a), we see that while the NSE obtained by optNSE is larger than
with optKGE, the difference is rather small. This indicates that by calibrating on KGE, we have
obtained only a slight deterioration in overall performance as measured by NSE. Further, although
there is a pronounced reduction in NSE from calibration to evaluation period, the reduction is
similar for both optNSE and optKGE.

However, the change in NSE tells us little that is diagnostically useful about the causes of this
‘deterioration’ in overall model performance. Of more interest, are the values obtained for the three
criterion components. The results for the calibration period are discussed first. Note that the
distribution of $r$ is almost identical with either optNSE or optKGE (Fig. 9b, filled symbols),
indicating that both of the criteria have achieved similar hydrograph match in terms of shape and
timing. However, for the other two components, optKGE has achieved considerably better results.

Fig. 9c shows that there is a strong tendency for underestimation of $a$ by optNSE (filled circle
symbols), due to which only 18 percent of the basins are within 10 percent of the ideal value at
unity, whereas for optKGE (filled triangle symbols) all of the basins are within 10 percent of the
ideal value. Similarly optKGE yields good results for $\beta$ (Fig. 9d), with all of the basins having a
bias of less than 10 percent, while for optNSE 16 percent of the basins have a bias of greater than
10 percent. In general, optKGE results in a $\beta$ value that is much closer to the ideal value at unity
than with optNSE. Thus, the use of optKGE has resulted in all of the basins having $a$ and $\beta$ close to
their ideal values of unity during calibration. This now explains why we get such a high correlation
between NSE and KGE in Fig. 7d; because both $\alpha$ and $\beta$ are now almost constant across the basins
(here close to unity), the equations for KGE and NSE both become approximately linear functions
of $r$, and in fact we tend towards the relationship $\text{NSE}(\theta_{\text{optKGE}}) = 2*\text{KGE}(\theta_{\text{optKGE}})-1$.

Next we examine what happens for the evaluation period. In general, we see that the statistical
distributions of the three components have changed. The cumulative distribution function of $r$ has
shifted to lower values in a consistent manner for both optNSE and optKGE (Fig. 9b), so that both
methods yield again very similar results for timing and shape. However, the optKGE calibrations
have retained a median value of $\alpha$ close to unity (the same as during calibration) while the overall
variability in the distribution has increased around the median value (Fig. 9c). This indicates that
the statistical tendency to provide good reproduction of flow variability persists into the evaluation
period, but there is an increase in the noise so that the distribution has become much wider. In
contrast, the optNSE results continue to show a systematic tendency to underestimate $\alpha$ (variability
of flows) during the evaluation period along with a considerable increase in random noise.

Similarly, the cumulative distribution function of $\beta$ obtained by both methods remains centred close
to its calibration value while showing an increase in the variability (Fig. 9d). The small shift in the
median value may be caused by the fact that there is approximately 5% less precipitation during the
evaluation period. Clearly, the KGE criterion has provided model calibrations that are statistically
more desirable during calibration while providing results that remain statistically more consistent on
the independent evaluation period.

An interesting observation is that in a few basins the paradoxical case occurs where all three of the
criterion components improve with optKGE, but the value of NSE decreases when compared to the
NSE obtained with optNSE (Table 2). The reason for this is the interplay between the terms $\alpha$ and $r$
in the NSE equation (illustrated nicely in Fig. 1). It is therefore actually (counter intuitively) possible for both $\alpha$ and $r$ to get closer to unity while NSE gets smaller. This is, of course, because optimization on NSE seeks to make $\alpha = r$, and therefore 'punishes' solutions for which $\alpha$ is close to the ideal value of unity, while $r$ will always be smaller than unity.

**Table 2 near here**

As discussed earlier, it is likely that optimization with NSE will yield results where $\alpha$ is close to $r$. Fig. 10a shows a comparison between $r$ and $\alpha$ obtained by the two optimization cases for all of the basins. In general, when optimizing with NSE, the value of $\alpha$ is indeed very similar to $r$, which means that the variability of flows is systematically underestimated (as shown above), and $\alpha$ approaches the ideal value of unity in only one of the 49 basins. In contrast, when optimizing with KGE, the value of $\alpha$ is close to the ideal value of unity for most of the basins.

Consequently, as expected from the theoretical discussion, systematically different results are obtained by optNSE and optKGE for the slopes of the regression lines (Fig. 10b), where the cases of regressing the simulated against the observed values ($k_o$, Eq. 14) and regressing the observed against the simulated values ($k_s$, Eq. 13) are distinguished. In general, when using optNSE the value of $k_s$ is close to the ideal value at unity, but $k_o$ is significantly smaller than one. In the case of optKGE both $k_s$ and $k_o$ are smaller than one, but the underestimation is not as large as for $k_o$ with optNSE. Note (from Eq. 12) that the only way that we can have both $k_s$ and $k_o$ equal to one is for $r$ to be equal to unity, which would only happen if the model and data were perfect.

**Fig. 10 near here**

Finally, we report briefly on the optimal parameter values obtained using optNSE and optKGE. Interestingly, even though the statistical properties of the streamflow hydrographs (as measured by $\alpha$ and $\beta$) did change significantly (Fig. 9), for many basins the parameter values did not change by
large amounts (compared to the feasible parameter range) when moving from optNSE to optKGE (Fig. 11). The correlation between the parameter values of optNSE and optKGE is at least 0.80 for all six of the parameters, and for three of the parameters the correlation is larger than 0.90. For the parameter \( K_1 \) the values are slightly smaller with optKGE, which has the effect of higher peaks and quicker recession of surface flow. Also the parameter \( K_3 \) decreases with optKGE, which has the effect of a less dampened base flow response. Given the function of these two parameters in the model structure, a reduction in the parameter values has the effect of increasing the value of the \( \alpha \) measure. In addition, we see an increase in the percolation parameter \( K_2 \), which results in more surface flow and less base flow, with the overall effect of increasing the value of \( \alpha \).

The function of the parameters \( Sl_{\max} \) and \( Beta \) in the model is mainly to control the partitioning of precipitation into runoff and evapotranspiration (thereby controlling the water balance), and as a consequence these parameters mainly affect the \( \beta \) measure. However, these parameters also affect the \( \alpha \) measure and parameter interaction between \( Sl_{\max} \) and \( Beta \) complicates the analysis. Given the function of these parameters in the model, the \( \beta \) measure should increase with a decrease in either \( Sl_{\max} \) and/or \( Beta \), but this is not obvious from Fig. 11, because a decrease in \( Sl_{\max} \) can be compensated by an increase in \( Beta \), and vice versa.

For the parameter \( S_2_{\phi n} \) no clear tendency of change is visible. Here it should be mentioned that there was no change in the parameter values in sixteen of the basins for which the parameter values were at their lower bounds (4 basins) and upper bounds (12 basins), respectively. Note that these 16 points also contribute to the rather high correlation observed.

On a visual, albeit subjective, basis a comparison of the parameter sets obtained by optNSE and optKGE reveals that in many of the basins the two parameter sets are almost indistinguishable, but
nevertheless the criterion components have changed. As an example, Fig. 12a displays a comparison of the parameters obtained by optNSE and optKGE for the Glan River. Apparently the parameter values are quite similar, although the $\alpha$ measure and to a lesser extent the $\beta$ measure have both improved when using optKGE (see Table 2). For many basins, the difference in each of the parameters was found to be only a small percentage of the overall feasible range (Fig. 12b); in 14 of the 49 basins, all of the six parameters have changed by less than 10%, and in only a few of the basins did two or more parameters change by a significant amount. For the latter, the changes may also (at least in part) be a consequence of parameter interactions; for example, there is a clear tendency for $K_2$ and $S_2$ to increase/decrease simultaneously, and this fact must, of course, also be considered when interpreting the scatter plots in Fig. 11.

4 Discussion

A decomposition of the NSE criterion shows that this measure of overall model performance can be represented in terms of three components, which measure the linear correlation, the bias and the variability of flow. By simple theoretical considerations, we can show that problems can arise in model calibrations that seek to optimize the value of NSE (or its related MSE). First, because the bias is normalized by the standard deviation of the observed flows, the relative importance of the bias term will vary across basins (and also across years), and for cases where the variability in the observed flows is high, the bias will have a low ‘weight’ in the computation of NSE. Second, there will be a tendency for the variability in the flows to be systematically underestimated, so that the ratio of the simulated and observed standard deviations of flows will tend to be equal to the correlation coefficient. As a consequence, the slope of the regression line (when regressing simulated against observed values) will be smaller than one, so that runoff peaks will tend to be
systematically underestimated. This finding may seem to contradict the general notion that optimization on NSE will improve simulation of runoff peaks. In fact NSE is generally found to be highly sensitive to the large runoff values, because of the (typically) larger model and data errors involved in the matching of such events, and this fact is separate from the general (theoretical and practical) tendency to underestimate the runoff peaks. Of course, when it is of interest to regress the observed against the simulated values then an optimization on NSE can yield desirable results, since in such a case the optimal slope of the regression line for maximizing NSE is equal to unity.

These theoretical considerations were all supported by the results of the modelling experiment. Of course, in such an experiment, not all solutions within the theoretical criteria space are possible because of constraints regarding the model structure, parameter ranges, and available data. However, it was found that the simple model was capable of achieving good solutions for the bias and the variability of flows with only slight decreases in the correlation coefficient. The optimization task therefore becomes one of specifying the objective function in such a way that it is capable of achieving such a solution as an optimal solution (i.e. simultaneously good solutions for bias, flow variability and correlation). Apparently, this was not the case with NSE, and we therefore formulated an alternative criterion (KGE) that is based on an equal weighting of the three components (correlation, bias, and variability measures). Of course the correlation will not, in general, reach its ideal value of unity, but an optimization on KGE resulted in the other two components being indeed close to their ideal values. Thus, the use of KGE instead of NSE for model calibration improved the bias and the variability measure considerably while only slightly decreasing the correlation.

The simulation results were also examined for an independent evaluation period. In general, the overall model performance and the individual components deteriorated in a statistical sense. It is at least partially likely that this is due to the rather short lengths of the calibration and evaluation
periods used in this study (five and three years, respectively). Further, it should be noted that this study has not accounted for either the uncertainty in the parameter values or the uncertainty in the computed statistics, which would require a more rigorous Bayesian approach. Nevertheless, the results clearly show that optimization using NSE tends to underestimate the variability of flows on the calibration period, and that this behaviour tends to persist into the evaluation period. Further, the bias in the calibration period is well constrained with KGE, but not with NSE, whereas in the evaluation period (with overall poorer bias) the results with NSE are only slightly inferior to KGE.

An interesting result is that for many basins the optimal parameter values changed by only small amounts (relative to the feasible range) when using KGE instead of NSE. In the KGE optimization there was a tendency to decrease the recession parameters of surface flow and base flow to simulate a flashier hydrograph, and thereby improve the value of the variability measure. Because of parameter interactions there was no clear tendency of a change in the parameters for the bias measure. In general, this suggests that the values of multiple criteria can be improved by making only small changes in the parameter values. This emphasizes the importance of the relative sensitivity of the criterion components to changes in the parameter values. On the one hand, this is a desirable effect during calibration, because we want to have measures that are actually sensitive to the parameter values, thereby theoretically increasing parameter identifiability. On the other hand, this raises important questions for parameter regionalization, because even a small ‘error’ in a parameter value could result in poor values of individual measures, thereby causing poor overall model performance.

The attempt to explain the relationships between changes in the parameters and values of the criterion components relates to the idea of diagnostic model evaluation, as proposed by Gupta et al. (2008) and tested by Yilmaz et al. (2008) and Herbst et al. (2009). The idea behind diagnostic model evaluation is to move beyond aggregate measures of model performance that are primarily...
529 statistical in meaning, towards the use of (multiple) measures and signature plots that are selected
530 for their ability to provide hydrological interpretation. Such an approach should improve our ability
531 to diagnose the causes of a problem and to make corrections at the appropriate level (i.e. model
532 structure or parameters). The theoretical development presented in this paper, shows one simple,
533 statistically founded approach to the development of a strategy for diagnostic evaluation and
534 calibration of a model. Clearly, the measures used in this study have some diagnostic value. The
535 bias and variability measures represent differences in matching of the means and the standard
536 deviations (the first two moments) of the probability distributions of the quantities being compared.
537 Their appearance in NSE and MSE indicates that these performance criteria give importance to
538 matching these two long-term statistics of the data. From a hydrological perspective, these statistics
539 relate to the properties of the flow duration curve, in which issues of timing and shape of the
540 dynamical characteristics of flow are largely ignored. These statistics will therefore be mainly
541 controlled by aspects of model structure and values of the parameters that determine the general
542 partitioning of precipitation into runoff, evapotranspiration and storage (i.e. overall water balance)
543 and, further, the general partitioning of runoff into fast and slow flow components (e.g. see Yilmaz
544 et al 2008). Meanwhile, all other differences between the statistical properties of the observed and
545 simulated flows such as timing of the peaks, and shapes of the rising limbs and the recedions of the
546 hydrograph (i.e. autocorrelation structures), are lumped into the (linear) correlation coefficient as an
547 aggregate measure. A logical next step would be to further decompose the correlation coefficient
548 into diagnostic components that represent different aspects of flow timing and shape (e.g.
549 autocorrelation structure). Further, a distinction between different states (modes) of the hydrological
550 response - such as driven and non-driven (see e.g. Boyle et al. 2000) – may also prove to be
551 sensible. Such considerations are left for future work.
Before entering into our concluding remarks, we should point out that the primary purpose of this study was not to design an improved measure of model performance, but to show clearly that there are systematic problems inherent with any optimization that is based on mean squared errors (such as NSE). The alternative criterion KGE was simply used for illustration purposes. An optimization on KGE is equivalent to selecting a point from the three-dimensional Pareto front with the minimal distance from the ideal point. Many different alternative criteria would also be sensible, but ultimately it has to be understood that each single measure of model performance has its own peculiarities and trade-offs between components. In the case of KGE probably the most problematic characteristic is that the slope of the regression lines will tend to be smaller than one, albeit not as strongly as with NSE (when regressing simulated against observed values). Because of the simple design of the KGE criterion it is straightforward to understand the trade-offs between the correlation, the bias and the variability measure. These trade-offs are more complex in the case of NSE.

If single measures of model performance are used we deem it to be imperative to clearly know the limitations of the selected criterion. It then will depend upon the type of application whether these limitations are of concern or not. The decomposition presented here highlights the fact that identical values of the NSE criterion are not necessarily indistinguishable - as is commonly (and erroneously) assumed in the literature in arguments relating to equifinality (Beven and Binley 1992, Beven and Freer 2001) - since the criterion components may be quite different. Thus, when evaluating or reporting results based on calibration with NSE, information about the correlation, bias, and variability of flows should also be given (interestingly, this was already proposed by Legates and McCabe (1999), although they did not discuss the interrelation between NSE and its three components). Ultimately the decision to accept or reject a model must be made by an expert hydrologist, where such a decision is best based in a multiple-criteria framework. To this end, an
analysis of the components that constitute the overall model performance can significantly enhance our understanding of model behaviour and provide insights helpful for diagnosing differences between models, basins and time periods within a hydrological context.

5  Summary and conclusions

In this study a decomposition of the widely used Nash-Sutcliffe efficiency (NSE) was applied to analyse the different components that constitute NSE (and hence MSE). We present theoretical considerations that serve to highlight problems associated with the NSE criterion. The results of a case study, where a simple precipitation-runoff model was applied in several basins, support the theoretical findings. For comparison we show how an alternative measure of model performance (KGE) can overcome the problems associated with NSE.

In summary, the main conclusions of this study are:

- The mean squared error and its related NSE criterion consists of three components, representing the correlation, the bias and a measure of variability. The decomposition shows that in order to maximize NSE the variability has to be underestimated. Further, the bias is scaled by the standard deviation in the observed values, which complicates a comparison between basins.

- Given that NSE consists of three components, an alternative model performance criterion KGE is easily formulated by computing the Euclidian distance of the three components from the ideal point, which is equivalent to selecting a point from the three-dimensional Pareto front. Such an alternative criterion avoids the problems associated with NSE (but also introduces new problems).
• The slopes of the regression lines are directly related with the three components. NSE is suitable if the interest is in regressing the observed against the simulated values, but less suitable for regressing the simulated against the observed values. This means that if NSE is used in optimization, then runoff peaks will tend to be underestimated. The same applies for KGE, but the underestimation will not be as severe.

• After optimization, the component representing the linear correlation dominates the model performance criterion for both NSE and KGE. For non-optimal parameters sets any of the three components can be dominant in NSE or KGE.

• Even with a simple precipitation-runoff model it is possible to obtain runoff simulations where the mean and variability of flows are matched well, and the linear correlation is still high. However, this applies only for optimization with KGE, since NSE does not consider such a solution as ‘good’.

• The optimal parameter values may, in practice, only change by small amounts when using KGE instead of NSE as the objective function for optimization (as in our example). This emphasizes the importance of considering the sensitivity of the three components to perturbations in the parameter values.

This study reinforces the argument that model calibration is a multi-objective problem (Gupta et al 1998), and shows that a decomposition of the calibration criterion into components can help to greatly enhance our understanding of the overall model performance (and, by extension, the differences in model performance between model structures, basins and time periods). To compute these components is a straightforward task and should be included in any evaluation of model simulations. Ultimately, such an approach may help in the design of diagnostically powerful evaluation strategies that properly support the identification of hydrologically consistent models.
6 Acknowledgements

Funding was provided for Harald Kling as an Erwin Schrödinger Scholarship (grant number J2719-N10) by FWF, Vienna, Austria. Partial support for Hoshin Gupta and Koray Yilmaz was provided by the Hydrology Laboratory of the National Weather Service (Grant NA04NWS4620012) and by SAHRA (Center for Sustainability of semi-Arid Hydrology and Riparian Areas) under the STC program of the National Science Foundation (agreement EAR 9876800). Partial support for Koray Yilmaz was also provided by NASA’s Applied Sciences Program (Stephen Ambrose) and PMM (Ramesh Kakar) of NASA Headquarters.

7 References


Rode M, Suhr U, Wriedt G. 2007. Multi-objective calibration of a river water quality model - 
Information content of calibration data. Ecological Modelling 204: 129-142

alpine regions in groundwater flow models. Advances in Geosciences 5: 19-23


Thornthwaite CW, Mather JR. 1957. Instructions and tables for computing potential evaporation 
and the water balance. Publications in Climatology 10(3): 311pp

Tolson BA, Shoemaker CA. 2007. Dynamically dimensioned search algorithm for computationally 
efficient watershed model calibration. Water Resources Research 43: 16 pp

van Griensven A, Bauwens W. 2003. Multiobjective autocalibration for semidistributed water 
quality models. Water Resources Research 39(12): 9pp

Wang G, Xia J, Chen J. 2009. Quantification of effects of climate variations and human activities 
on runoff by a monthly water balance model: A case study of the Chaobai River basin in northern 
China. Water Resources Research 45: 12pp

Weglarczyk S. 1998. The interdependence and applicability of some statistical quality measures for 
hydrological models. Journal of Hydrology 206: 98-103

evaluation: application to the NWS distributed hydrologic model. Water Resources Research 44: 
18pp

Young AR. 2006. Stream flow simulation within UK ungauged catchments using a daily rainfall-
runoff model. Journal of Hydrology 320: 155-172
Table 1: Parameters of the model. Parameters in brackets were not calibrated.

<table>
<thead>
<tr>
<th>parameter</th>
<th>units</th>
<th>feasible range</th>
<th>description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$SI_{max}$</td>
<td>mm</td>
<td>50 - 700</td>
<td>soil storage capacity</td>
</tr>
<tr>
<td>Beta</td>
<td>/</td>
<td>0.1 - 25</td>
<td>exponent for computing runoff generation</td>
</tr>
<tr>
<td>$(SI_{crit})$</td>
<td>/</td>
<td>(0.6)</td>
<td>critical soil moisture for actual evapotranspiration</td>
</tr>
<tr>
<td>$K1$</td>
<td>h</td>
<td>10 - 500</td>
<td>recession coefficient for surface flow</td>
</tr>
<tr>
<td>$K2$</td>
<td>h</td>
<td>10 - 1000</td>
<td>recession coefficient for percolation</td>
</tr>
<tr>
<td>$S2_{crit}$</td>
<td>mm</td>
<td>0 - 15</td>
<td>outlet height for surface flow</td>
</tr>
<tr>
<td>$K3$</td>
<td>h</td>
<td>500 - 10000</td>
<td>recession coefficient for base flow</td>
</tr>
<tr>
<td>$(K4)$</td>
<td>h</td>
<td>(0 - 10)</td>
<td>recession coefficient for distributed routing</td>
</tr>
</tbody>
</table>

Table 2: ‘Paradoxical’ examples for NSE and components in three basins (results for the calibration period). All three components ($r$, $\alpha$, $\beta$) improve but the overall model performance measured by NSE decreases with the parameter set obtained by optKGE.

<table>
<thead>
<tr>
<th>basin</th>
<th>method</th>
<th>NSE []</th>
<th>KGE []</th>
<th>$r$ []</th>
<th>$\alpha$ []</th>
<th>$\beta$ []</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zaya River</td>
<td>optNSE</td>
<td>0.484</td>
<td>0.685</td>
<td>0.714</td>
<td>0.871</td>
<td>1.019</td>
</tr>
<tr>
<td>optKGE</td>
<td></td>
<td>0.452</td>
<td>0.732</td>
<td>0.733</td>
<td>1.026</td>
<td>1.001</td>
</tr>
<tr>
<td>Pitten River</td>
<td>optNSE</td>
<td>0.742</td>
<td>0.828</td>
<td>0.863</td>
<td>0.899</td>
<td>1.028</td>
</tr>
<tr>
<td>optKGE</td>
<td></td>
<td>0.730</td>
<td>0.865</td>
<td>0.866</td>
<td>1.004</td>
<td>1.016</td>
</tr>
<tr>
<td>Glan River</td>
<td>optNSE</td>
<td>0.786</td>
<td>0.855</td>
<td>0.888</td>
<td>0.912</td>
<td>1.028</td>
</tr>
<tr>
<td>optKGE</td>
<td></td>
<td>0.776</td>
<td>0.888</td>
<td>0.889</td>
<td>1.002</td>
<td>1.007</td>
</tr>
</tbody>
</table>
Figure captions:

Fig. 1: Relationship of NSE with $\alpha$ and $r$ ($\beta_n$ is assumed to be zero): (a) theoretical relationship, (b) illustrative example obtained by random parameter sampling with a hydrological model (Leaf River, Mississippi, USA, 1924 km$^2$, 11 years daily data, HyMod model; only those points where $\beta_n^2 < 0.01$ are displayed). Contour lines indicate values for NSE. See colour version of this figure online.

Fig. 2: Example for three-dimensional Pareto front of $r$, $\alpha$ and $\beta$. $ED$ is the Euclidian distance between the ‘optimal’ point and the ideal point, where all three measures are 1.0. Glen River, Austria, 432 km$^2$, 5 years daily data, HBV model variant, random parameter sampling.

Fig. 3: Typical scatter plots depicting simulated and observed runoff ($r = 0.86$ and $\alpha = 0.90$) and fitted regression lines: (a) regression against simulated runoff ($k_s = 0.96$) and (b) regression against observed runoff ($k_o = 0.77$). Pitten River, Austria, 277 km$^2$, 5 years daily data, HBV model variant, parameters optimized on NSE. Note, that in (a) and (b) the identical data points are plotted, but the axes are flipped.

Fig. 4: Map showing locations of the 49 Austrian basins used in this study. Also depicted are the 49 gauges and 222 precipitation stations.

Fig. 5: Relationship between index of evaporation and index of wetness for the 49 Austrian basins. The index of wetness is computed as the ratio between precipitation ($P$) and potential evapotranspiration ($ET_p$). The index of evaporation is computed as the ratio between actual evapotranspiration ($ET_a$) and $ET_p$. Data represent long-term means from the period 1961 to 1990 and are taken from Hydrological Atlas of Austria (BMLFUW 2007).

Fig. 6: Conceptual model structure (the snow module is not shown). Parameters in brackets are not calibrated.
Fig. 7: Scatter plots of overall model performance: cal = calibration period, eval = evaluation period. Note that in (a) two points are located outside the plotting range because of negative NSE values in the evaluation period.

Fig. 8: Stacked area plots showing the relative contribution of the components for NSE and KGE in the calibration period: (a) optNSE in 49 basins, (b) optKGE in 49 basins, (c) and (d) random parameter sampling in the Glan River basin.

Fig. 9: Cumulative distribution functions for NSE, $r$, $\alpha$, and $\beta$ as obtained with optNSE and optKGE in the calibration and evaluation periods.

Fig. 10: Relationship between (a) $r$ and $\alpha$ and (b) the slope of the regression lines $k$ and $k_o$.

Fig. 11: Scatter plots of optimal parameters obtained by optNSE and optKGE. Parameter values are normalized by the feasible parameter range (Table 1); the parameters $Beta$, $K1$, $K2$ and $K3$ are log-transformed before normalization.

Fig. 12: Comparison of the parameter sets obtained by optNSE and optKGE: (a) normalized parameter values of $\theta_{optNSE}$ and $\theta_{optKGE}$ in the Glan River basin, (b) difference in the normalized parameter values (computed as $\theta_{optKGE} - \theta_{optNSE}$), displayed for all basins.
Figure 1
Figure 3

(a) regression against simulated runoff

(b) regression against observed runoff

$k_s = 0.96$

$k_o = 0.77$
Figure 5

[Graph showing the relationship between Index of wetness (P/ETp) and Index of evaporation (E/Ta/ETp) for different locations, including Glan, Pitten, and Zaya.]
Figure 6

Act. ET

Rainfall + Snowmelt

act, ET

Pot. ET

(S1<sub>crit</sub>)

S1<sub>max</sub>

Beta

surface flow reservoir

K1

K2

base flow reservoir

K3

upstream

Runoff

Routing

K4

Runoff
Figure 7
Figure 8

49 basins, optNSE

NSE rank of basin

NSE [1, 0]

5 10 15 20 25 30 35 40 45

Glän River, 5000 random parameter samples

NSE rank of sample

Glän River, 5000 random parameter samples

KGE rank of basin

KGE [1, 0]

5 10 15 20 26 30 36 41

KGE rank of sample

KGE [1, 0]
Figure 9
Figure 12

(a) Glan River basin normalized parameter value [f] vs. Smax, Beta, K1, K2, S2opt, K3

(b) Difference in normalized parameter value [f] for all other basins and Glan River basin.

SUMMARY

The mean squared error (MSE) and the related normalization, the Nash-Sutcliffe efficiency (NSE), are the two criteria most widely used for calibration and evaluation of hydrological models with observed data. Here, we present a diagnostically interesting decomposition of NSE (and hence MSE), which facilitates analysis of the relative importance of its different components in the context of hydrological modelling, and show how model calibration problems can arise due to interactions among these components. The analysis is illustrated by calibrating a simple conceptual precipitation-runoff model to daily data for a number of Austrian basins having a broad range of hydro-meteorological characteristics. Evaluation of the results clearly demonstrates the problems that can be associated with any calibration based on the NSE (or MSE) criterion. While we propose and test an alternative criterion that can help to reduce model calibration problems, the primary purpose of this study is not to present an improved measure of model performance. Instead, we seek to show that there are systematic problems inherent with any optimization based on formulations related to the MSE. The analysis and results have implications to the manner in which we calibrate and evaluate environmental models; we discuss these and suggest possible ways forward that may move us towards an improved and diagnostically meaningful approach to model performance evaluation and identification.