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

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,

Please find enclosed our manuscript entitled "Decomposition of the Mean Squared Error & NSE Performance Criteria: Implications for Improving Hydrological Modelling" submitted for publication in the Journal of Hydrology, authored by Hoshin V. Gupta, Harald Kling, Koray K. Yilmaz, Guillermo F. Martinez-Baquero.

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

Gupta, Kling, Yilmaz, Martinez-Baquero 2009, submitted to Journal of Hydrology, version 1.0

1 **Title:**

2 Decomposition of the Mean Squared Error & NSE Performance Criteria: Implications for

3 Improving Hydrological Modelling

4

5 **Authors:**

6 Hoshin V. Gupta[1], Harald Kling[1], Koray K. Yilmaz[1,2], Guillermo F. Martinez-Baquero[1]

7

8 **Affiliations:**

9 1: Department of Hydrology and Water Resources, The University of Arizona, 1133 E North

10 Campus Dr, Tucson, AZ 85721, USA

11 2: now at Earth System Science Interdisciplinary Center, University of Maryland, College Park,

12 MD 20742, USA and at NASA Goddard Space Flight Center, Laboratory for Atmospheres,

13 Greenbelt, MD 20771, USA

14

15 **Corresponding author:**

16 Harald Kling

17 e-mail: harald.kling@boku.ac.at

18 Phone: +1 520 626 9712

19 Fax: +1 520 621 1422

20

21   **Abstract:**

22   The mean squared error (MSE) and the related normalization, the Nash-Sutcliffe efficiency (NSE),

23   are the two criteria most widely used for calibration and evaluation of hydrological models with

24   observed data. Here, we present a diagnostically interesting decomposition of NSE (and hence

25   MSE), which facilitates analysis of the relative importance of its different components in the

26   context of hydrological modelling, and show how model calibration problems can arise due to

27   interactions among these components. The analysis is illustrated by calibrating a simple conceptual

28   precipitation-runoff model to daily data for a number of Austrian basins having a broad range of

29   hydro-meteorological characteristics. Evaluation of the results clearly demonstrates the problems

30   that can be associated with any calibration based on the NSE (or MSE) criterion. While we propose

31   and test an alternative criterion that can help to reduce model calibration problems, the primary

32   purpose of this study is not to present an improved measure of model performance. Instead, we seek

33   to show that there are systematic problems inherent with any optimization based on formulations

34   related to the MSE. The analysis and results have implications to the manner in which we calibrate

35   and evaluate environmental models; we discuss these and suggest possible ways forward that may

36   move us towards an improved and diagnostically meaningful approach to model performance

37   evaluation and identification.

38   **Keywords:**

39   mean squared error; Nash-Sutcliffe efficiency; model performance evaluation; calibration; multiple

40   criteria; hydrologic modelling; criteria decomposition; diagnostic analysis

41

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:

$$MSE = \frac{1}{n} \cdot \sum_{t=1}^{n} \left(x_{s,t} - x_{o,t} \right)^2 \qquad \text{Eq. 1}$$

$$NSE = 1 - \frac{\sum_{t=1}^{n} \left(x_{s,t} - x_{o,t} \right)^2}{\sum_{t=1}^{n} \left(x_{o,t} - \mu_o \right)^2} = 1 - \frac{MSE}{\sigma_o^2} \qquad \text{Eq. 2}$$

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 μ_o and σ_o are the mean and standard deviation of the observed values. In optimization MSE is subject to minimization and NSE is subject to maximization.

57 While the NSE criterion may be a convenient and popular (albeit gross) indicator of model skill,

58 there has been a long and vivid discussion about the suitability of NSE (McCuen and Snyder 1975,

59 Martinec and Rango 1989, Legates and McCabe 1999, Krause et al. 2005, McCuen et al. 2006,

60 Schaefli and Gupta 2007) and several authors have proposed modifications - e.g. Mathevet et al.

61 (2006) proposed a bounded version of NSE and Criss and Winston (2008) proposed a volumetric

62 efficiency to be used instead of NSE. One of the main concerns about NSE is its use of the observed

63 mean as baseline, which can lead to overestimation of model skill for highly seasonal variables such

64 as runoff in snowmelt dominated basins. A comparison of NSE across basins with different

65 seasonality (as is often reported in the literature) should therefore be interpreted with caution. For

66 such situations, various authors have recommended the use of the seasonal or climatological mean

67 as a baseline model (Garrick et al. 1978, Murphy 1988, Martinec and Rango 1989, Legates and

68 McCabe 1999, Schaefli and Gupta 2007).

69 It is now generally accepted that the calibration of hydrological models should be approached as a

70 multi-objective problem (Gupta et al. 1998). Within a multiple-criteria framework, the MSE and

71 NSE criteria continue to be commonly used, because they can be computed separately for (1)

72 different types of observations (e.g. runoff and snow observations; Bergström et al. 2002), (2)

73 different locations (e.g. runoff at multiple gauges; Madsen 2003), or (3) different subsets of the

74 same observation (e.g. rising and falling limb of the hydrograph; Boyle et al. 2000). More generally,

75 however, different types of model performance criteria - such as NSE, coefficient of correlation,

76 bias, etc. - can be computed from multiple variables and/or at multiple sites (see Anderton et al.

77 2002, Beldring 2002, Rojanschi et al. 2005, Cao et al. 2006, and others).

78 When handled in this manner, the model calibration problem can be treated as a full multiple-

79 criteria optimization problem resulting in a 'Pareto set' of non-dominated solutions (Gupta et al

80 1998), or reduced to a related single-criterion optimization problem by combining the different

81   (weighted) criteria into one overall objective function. Numerous examples of the latter approach

82   exist in the literature where NSE or MSE appear in an overall objective function (e.g. Lindström

83   1997, Bergström et al. 2002, Madsen 2003, van Griensven and Bauwens 2003, Parajka et al. 2005,

84   Young 2006, Rode et al. 2007, Marce et al. 2008, Wang et al. 2009), because it conveniently

85   enables the application of efficient single-criterion automated search algorithms, such as SCE

86   (Shuffled Complex Evolution, Duan et al. 1992) or DDS (Dynamically Dimensioned Search,

87   Tolson and Shoemaker 2007).

88   When using multiple criteria in evaluation, it has to be considered that some of these criteria are

89   mathematically related, which is not always recognized (Weglarczyk 1998). For example, it is

90   possible to decompose the NSE criterion into separate components, as shown by Murphy (1988)

91   and Weglarczyk (1998), which facilitates a better understanding of how different criteria are

92   interrelated and thereby enable more insight into what is causing a particular model performance to

93   be 'good' or 'bad'. Equally important, the decomposition can provide insight into possible trade-

94   offs between the different components.

95   In this paper we present a diagnostically interesting decomposition of NSE (and hence MSE), which

96   facilitates analysis of the relative importance of different components in the context of hydrological

97   modelling, and show how model calibration problems can arise due to interactions among these

98   components. Based on this analysis, we propose and test alternative criteria that can help to avoid

99   these problems. The analysis is illustrated by calibrating a simple precipitation-runoff model to

100   daily data for a number of Austrian basins having a broad range of hydro-meteorological

101   characteristics, and evaluating the results on both the calibration and an independent 'evaluation'

102   period. The results clearly demonstrate the problems that can be associated with any calibration

103   based on the NSE (or MSE) criterion. The analysis and results have interesting implications to the

104 manner in which we calibrate and evaluate environmental models; we discuss these and some

105 possible ways forward in the discussion and conclusions sections.

2   Decomposition of model performance criteria

2.1   Decomposition of NSE

108 A decomposition of criteria based on mean squared errors reveals that there are three distinctive

109 components, represented by the correlation, the conditional bias, and the unconditional bias, as

110 evident in Eq. 3 which shows a decomposition of NSE (Murphy 1988, Weglarczyk 1998).

$$NSE = A - B - C \qquad \text{Eq. 3}$$

111 with: $A = r^2$

112 $$B = \left[r - \left(\sigma_s / \sigma_o \right) \right]^2$$

113 $$C = \left[\left(\mu_s - \mu_o \right) / \sigma_o \right]^2$$

114 where r is the linear correlation coefficient between x_s and x_o, and (μ_s, σ_s) and (μ_o, σ_o) represent the

115 first two statistical moments (means and standard deviations) of x_s and x_o respectively. The quantity

116 A measures the strength of the linear relationship between the simulated and observed values, B

117 measures the conditional bias, and C measures the unconditional bias (Murphy 1988).

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

$$NSE = 2 \cdot \alpha \cdot r - \alpha^2 - \beta_n^2 \qquad \text{Eq. 4}$$

119 with: $\alpha = \sigma_s / \sigma_o$

120 $$\beta_n = \left(\mu_s - \mu_o\right)/\sigma_o$$

121 where the quantity α is a measure of relative variability in the simulated and observed values, and

122 β_n is the bias normalized by the standard deviation in the observed values (note that $\beta_n = \text{sqrt}(C)$).

123 Eq. 4 shows that NSE is composed of three components, two of which relate to the ability of the

124 model to reproduce the *first and second moments of the distribution of the observations* (i.e. mean

125 and standard deviation), while the third relates to the *ability of the model to reproduce timing and*

126 *shap*e as measured by the correlation coefficient. The ideal values for the three components are

127 $r = 1$, $\alpha = 1$, and $\beta_n = 0$. From a hydrological perspective, 'good' values for each of these three

128 components are highly desirable, since in general we aim at matching the overall volume of flow,

129 the spread of flows (e.g. flow duration curve), and the timing and shape of (for example) the

130 hydrograph (Yilmaz et al. 2008). It is clear, therefore, that optimizing NSE is essentially a search

131 for a balanced solution among the three components, which is similar to the multiple-criteria

132 approach of computing an overall (weighted) objective function from several different criteria as

133 discussed in the introduction.

134 However, in using NSE we must be concerned with two facts. First, the bias (μ_s - μ_o) component

135 appears in a normalized form, scaled by the standard deviation in the observed flows. This means

136 that in basins with high runoff variability the bias component will tend to have a smaller

137 contribution (and therefore impact) in the computation and optimization of NSE, possibly leading to

138 model simulations having large volume balance errors. In a multiple-criteria sense, this is

139 equivalent to using a weighted objective function with a low weight applied to the bias component.

140 Second, and equally serious, the quantity α appears twice in Eq. 4, exhibiting an interesting (and

141 problematic) interplay with the linear correlation coefficient r. It is easy to show, by taking the first

142 derivative of NSE (in Eq. 4) with respect to α, that the maximum value of NSE is obtained when

143    $\alpha = r$. And, since r will always be smaller than unity, this means that in maximizing NSE we will

144   tend to select a value for α that *underestimates the variability in the flows* (more precisely, we will

145   favour models/parameter sets that generate simulated flows that underestimate the variability).

146   Taking these two facts together, we note that when $\beta_n = 0$ and $\alpha = r$, then the NSE is equivalent to

147   r^2, which is the well-known coefficient of determination. Therefore, r^2 can be interpreted as a

148   maximum (potential) value for NSE if the other two components are able to achieve their 'optimal'

149   values.

150   Fig. 1 illustrates the relationship of NSE with r and α, while assuming that β_n is zero (β_n is only an

151   additive term, anyway). For a given r the 'optimal' α for maximizing NSE lies on the 1:1 line,

152   although the ideal value of α is on a horizontal line at 1.0. This theoretical relationship is illustrated

153   in Fig. 1a. Of course, not all combinations of r and α may be possible with a hydrological model

154   due to restrictions imposed by the model structure, feasible parameter values and input-output data.

155   However, Fig. 1b shows a real example in which random sampling of the parameter space actually

156   seems to cover a large portion of the theoretical criteria space. Since the model used here (HyMod

157   model, Boyle 2000) is a simple, but representative, example of watershed models in common use,

158   the problematic interplay between α and r is likely to be of importance for any type of hydrological

159   model that is optimized with NSE.

160   **Fig. 1 near here**

161   Further, the same exact problems will arise when using MSE as a model calibration criterion. We

162   can substitute Eq. 4 into Eq. 2, and thereby obtain Eq. 5 which shows the related decomposition of

163   the MSE criterion, consisting (again) of three error terms, but here all three of them are additive.

$$MSE = 2 \cdot \sigma_s \cdot \sigma_o \cdot (1-r) + (\sigma_s - \sigma_o)^2 + (\mu_s - \mu_o)^2 \qquad \text{Eq. 5}$$

164 From Eq. 3, Eq. 4 and Eq. 5 it should be immediately obvious that many different combinations of

165 the three components can result in the same overall value for MSE or NSE, respectively, potentially

166 leading to considerable ambiguity in the comparative evaluation of alternative model hypotheses.

167 The relative contribution of each of these components to the overall MSE can be computed as:

$$f_i = F_i / \sum_{j=1}^{3} F_j \qquad \text{Eq. 6}$$

168 with: $\quad F_1 = 2 \cdot \sigma_s \cdot \sigma_o \cdot (1 - r)$

169 $\qquad F_2 = (\sigma_s - \sigma_o)^2$

170 $\qquad F_3 = (\mu_s - \mu_o)^2$

171   2.2   Alternative model performance criteria

172 As discussed above, a peculiar feature of the NSE criterion is the problematic interplay between α

173 and r, which is likely to result in an underestimation of the variability in the flows. One way to

174 overcome this is by inflating the observed variability as indicated by Eq. 7, while at the same time

175 preserving the mean of the observations and their linear correlation with the simulations. Using Eq.

176 7 with Eq. 4 results in Eq. 8, which represents a 'corrected' version of NSE:

$$x_{o,t}^* = c \cdot (x_{o,t} - \mu_o) + \mu_o \qquad \text{Eq. 7}$$

$$NSE_{cor} = \frac{1}{c} \cdot 2 \cdot \alpha \cdot r - \frac{1}{c^2} \cdot \alpha^2 - \frac{1}{c^2} \cdot \beta_n^2 \qquad \text{Eq. 8}$$

177   where c is correction factor to inflate the variability in the observed flows. It can be easily shown

178   that if c is set equal to $1/r$, it will assure that a value of $\alpha = 1$ will now maximize NSE_{cor} (as opposed

179   to $\alpha = r$ maximizing NSE).

180   Alternatively, instead of trying to come up with a 'corrected' NSE criterion, since MSE and NSE

181   can be decomposed into three components, the whole calibration problem can instead be viewed

182   from the multi-objective perspective, by focusing on the correlation, variability error and bias error

183   as separate criteria to be optimized. In doing this, it makes sense to enable a better hydrological

184   interpretation of the bias component by using the ratio of the means of the simulated and observed

185   flows (β) for this further analysis - as opposed to using β_n. With this formulation, using β instead of

186   β_n, all three of the components now have their ideal value at unity.

187   Fig. 2 shows an example for the trade-off between the three components for a simple hydrological

188   model using random parameter sampling. The plot shows a distinctive Pareto front in the three-

189   dimensional criteria space. If it is desired to select a compromise solution from the Pareto front, one

190   possible approach is to compute for all points the Euclidian distance from the ideal point and then to

191   subsequently select the point having the shortest distance (Eq. 9). Since all three of the components

192   are dimensionless numbers, we are able to obtain a reasonable solution for the Euclidian distance in

193   the un-transformed criteria space. Alternatively, a re-scaling of the axes in the criteria space is

194   easily obtained via Eq. 10. In this paper, we will only explore the use of the KGE criterion (Eq. 9),

195   which is equivalent to setting all three scaling factors of Eq. 10 to unity.

196   **Fig. 2 near here**

$$KGE = 1 - ED \qquad \text{Eq. 9}$$

$$KGE_s = 1 - ED_s \qquad \text{Eq. 10}$$

197   with: $ED = \sqrt{(r-1)^2 + (\alpha-1)^2 + (\beta-1)^2}$

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

199    $\beta = \mu_s / \mu_o$

200   where ED is the Euclidian distance from the ideal point, ED_s is the Euclidian distance from the ideal

201   point in the scaled space, β is the ratio between the mean simulated and mean observed flows, i.e. β

202   represents the bias; s_r, s_α and s_β are scaling factors that can be used to re-scale the criteria space

203   before computing the Euclidian distance from the ideal point, i.e. s_r, s_α and s_β can be used for

204   adjusting the emphasis on different components.

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

$$g_i = G_i / \sum_{j=1}^{3} G_j \qquad \text{Eq. 11}$$

206   with: $G_1 = (r-1)^2$

207    $G_2 = (\alpha-1)^2$

208    $G_3 = (\beta-1)^2$

2.3   Notes on regression lines

210   As is well known, the slope of the regression lines and the coefficient of correlation are related (Eq.

211   12 to Eq. 14). Since different 'optimal' values for α are obtained by the NSE and KGE criteria, this

212   also leads to implications for the regression lines.

$$r^2 = k_s \cdot k_o \qquad \text{Eq. 12}$$

$$k_s = \frac{Cov_{so}}{\sigma_s^2} = \frac{r}{\alpha} \qquad \text{Eq. 13}$$

$$k_o = \frac{Cov_{so}}{\sigma_o^2} = r \cdot \alpha \qquad \text{Eq. 14}$$

213   with: $\qquad r = \dfrac{Cov_{so}}{\sigma_s \cdot \sigma_o}$

214   where Cov_{so} is the covariance between the simulated and observed values, k_s is the slope of the

215   regression line when regressing the observed against the simulated values, and k_o is the slope of the

216   regression line when regressing the simulated against the observed values.

217   Murphy (1988) has already noted that for NSE the conditional bias term B in Eq. 3 will vanish only

218   if the slope of the regression line k_s is equal to unity (i.e. regressing the observed against the

219   simulated values), which is desirable in the context of the 'verification' of forecasts. This means

220   that for a given forecast (simulated value), the expected value of the observed value lies on the 1:1

221   line (assuming a Gaussian distribution). As discussed before, the optimal value of α that maximizes

222   NSE is given by a model simulation for which α is equal to r. As evident in Eq. 13 this results in

223     $k_s = 1$, but at the same time this also implies that $k_o = r^2$ (Eq. 14). Because r^2 will always be smaller

224   than unity, this means that we will, in general, tend to underestimate the slope of the regression line

225   when regressing the simulated against the observed values. The tendency will be for high values

226   (peak flows) to be underestimated and for low values (recessions) to be overestimated in the

227   simulation.

228   In brief, for maximizing NSE the optimal values for k_s and k_o are unity and r^2, respectively. In the

229   case of KGE, the optimal value for α is at unity, which means that for maximizing KGE the optimal

230   values for both k_s and k_o are equal to r. Again, since r is smaller than unity we will tend to

231   underestimate high values and overestimate low values.

232   In considering this, it should be noted that both approaches for computing the regression lines

233   (regressing observed against simulated values, or vice versa) are valid, but have different

234   interpretations. In the context of runoff simulations, when using k_s we are basing the evaluation on

235   the expected error in simulation of the observed runoff being zero for a given simulated runoff,

236   which is a sensible approach when making runoff forecasts under 'normal' conditions. However, if

237   we are interested in the 'unusual' runoff conditions - such as runoff peaks - then a more sensible

238   approach would be to use k_o, where we are interested in the question, "If a flood occurs, can we

239   forecast (simulate) it?", whereas in the case of k_s such a runoff peak is 'averaged out'. Fig. 3

240   illustrates this with typical scatter plots for runoff simulation. In this example, k_s is close to unity,

241   suggesting unbiased forecasts (Fig. 3a), and at the highest simulated flows of around 10 m^3/s the

242   small number of observed flows (runoff peaks) that are well above the regression line are 'averaged

243   out' by the larger number of observed flows that occur slightly below the regression line. However,

244   it is clear that whenever a runoff peak above 10 m^3/s occurs, there is a clear tendency for

245   underestimation in the simulation (Fig. 3b).

246   These problems arise because the distribution of runoff is usually highly skewed. If k_o is of higher

247   interest, then the use of NSE may cause problems, since the simulated runoff will tend to

248   underestimate the peak flows. In the case of the KGE criterion, we will also have a tendency

249   towards underestimation, but not as severe as with the NSE. Note that for extreme low-flows,

250   similar considerations as for the runoff peaks apply (but here we will tend to overestimate the low-

251   flow).

13

252 **Fig. 3 near here**

3   Case Study

254   To examine and illustrate the implications of the theoretical considerations presented above we

255   applied a simple conceptual precipitation-runoff model to several basins. Using NSE (Eq. 2) and

256   KGE (Eq. 9) as model performance criteria, two different sets of parameters were obtained for each

257   basin by calibration against observed runoff data. For each parameter set we compare the overall

258   model performance as evaluated by the NSE and KGE criteria and, in addition, conduct a detailed

259   analysis of the criterion components. Further, we also examine the model performance on an

260   independent 'evaluation' period.

3.1   Study area

262   For this study we used the forty-nine mesoscale Austrian basins (Fig. 4) used in the regionalization

263   study reported by Kling and Gupta (2009). All are pre-alpine or lowland basins where snowmelt

264   does not dominate runoff generation. They vary in size from 112.9 km^2 to 689.4 km^2, with a median

265   size of 287.3 km^2, and a mean elevation range from 232 m to 952 m above sea level. The basins

266   represent a wide range of physiographic and meteorological properties, with the most important

267   land-use types being forest, grassland and agriculture. According to the Hydrological Atlas of

268   Austria (BMLFUW 2007), the long-term mean annual precipitation in the basins ranges from 507 to

269   1929 mm, and the corresponding runoff ranges from 44 to 1387 mm, resulting in a large range of

270   runoff coefficients (from 9 to 72 percent). Thus, both wet and dry basins are included. Fig. 5 shows

271   a diagnostic plot where normalized actual evapotranspiration is plotted against normalized

272   precipitation (both variables are scaled by potential evapotranspiration); it indicates that most of the

273   basins are energy limited and only a few of the basins are water limited.

274 **Fig. 4 near here**

275 **Fig. 5 near here**

3.2 Data basis

277 We used observed daily data for the period September 1990 to August 2000; the first two years

278 were used as a warm-up period, the next five years for calibration, and the final three years for

279 independent evaluation. Observed catchment outlet runoff data were used for parameter calibration

280 in each of the basins. Precipitation inputs were based on daily data from 222 stations, regionalized

281 using the method of Thiessen-Polygons. Air temperature inputs were based on data from 98

282 stations, regionalized via linear regression with elevation. Potential evapotranspiration inputs were

283 based on monthly fields of potential evapotranspiration (Kling et al. 2007) with a spatial resolution

284 of 1x1 km. The monthly potential evapotranspiration data were disaggregated to daily time-steps by

285 using daily data from 21 indicator stations, where the daily potential evapotranspiration was

286 computed using the Thornthwaite-method (Thornthwaite and Mather 1957).

3.3 Hydrological model

288 A simple, conceptual, spatially distributed daily precipitation-runoff model similar to the HBV

289 model (Bergström 1995) was used; the model was previously applied to these same basins by Kling

290 and Gupta (2009). The model uses a 1x1 km^2 raster grid for spatial discretization of the basins.

291 However, for simplicity, the current study assumes uniform parameter fields. Inputs to the model

292 are precipitation, air temperature, and potential evapotranspiration. The model consists of a snow

293 module, soil moisture accounting, runoff separation into different components, and a routing

294 module. Snowfall is determined from precipitation data using a threshold temperature, and

295 snowmelt is computed with the temperature-index method (see e.g. Hock 2003). Rainfall and

296   snowmelt are input to the soil module, where runoff generation is computed via an exponential

297   formulation that accounts for current soil moisture conditions (see e.g. Bergström and Graham

298   1998). Actual evapotranspiration depletes the soil moisture store; the rate of actual

299   evapotranspiration depends on current soil moisture conditions and potential evapotranspiration.

300   Runoff is separated into fast (surface flow) and slow (base flow) components by two linear

301   reservoirs having different recession coefficients. A further linear reservoir is used to simulate

302   channel routing of the runoff. Fig. 6 shows the conceptual structure of the model (the snow module

303   is not shown). The model equations are presented in Kling and Gupta (2009). Table 1 lists the most

304   important parameters of the model.

305   To reduce dimensionality of the parameter calibration problem, some of the model parameters are

306   set to plausible values and are not further calibrated. This applies to snow parameters, because snow

307   is of limited importance in the basins of this study, and to the channel routing parameters, which are

308   of limited importance at the daily time-step (the values of Kling and Gupta (2009) are used). In

309   addition, the critical soil moisture for reducing actual evapotranspiration is set to a constant value.

310   The six remaining parameters were calibrated using the Shuffled Complex Evolution optimization

311   algorithm (SCE, Duan et al. 1992), using six complexes.

312   **Fig. 6 near here**

313   **Table 1 near here**

314   ## 3.4 Results

315   The optimization runs resulted in two parameter sets for each basin. Optimization using the

316   'optNSE' method results in parameter sets 'θ_{optNSE}' that yield optimal runoff simulations

317   maximizing NSE (Eq. 4), while optimization using the 'optKGE' method results in parameter sets

318   'θ_{optKGE}' that yield optimal runoff simulations maximizing KGE (Eq. 9). A standard method for

319   reporting model performance in precipitation-runoff modelling studies is to present scatter plots of

320   NSE between calibration and evaluation periods (see e.g. Merz and Blöschl 2004). Fig. 7 displays

321   such a scatter plot; as expected, for many basins the NSE deteriorates when going from the

322   calibration to the evaluation period (Fig. 7a). Similar results are obtained for KGE (Fig. 7b).

323   Now, there can be different reasons for deterioration of model performance on the evaluation

324   period. These include over-fitting of the parameters to the calibration period, non-stationarity

325   between the calibration and evaluation periods, lack of 'power' in the objective function, etc.

326   Instead of falsification and model rejection, which would be a logical conclusion from such a result,

327   it is common practice to simply report the deterioration in the model performance and then to move

328   on. In our case, we can report that when moving from calibration to evaluation period the median

329   NSE has decreased from 0.76 to 0.59 and the median KGE has decreased from 0.86 to 0.72, but

330   what hydrological meaning do these numbers have? Here, an analysis of the different components

331   that constitute the overall model performance enables us to learn much more about the model

332   behaviour, differences between the calibration and evaluation periods, and also differences between

333   basins.

334   Before analysing the criterion components it is interesting to note the relationship between NSE and

335   KGE. Fig. 7 shows that when optimizing on KGE (optKGE) there is a strong correlation between

336   the values obtained for the KGE and NSE criteria (Fig. 7d). However, when optimizing on NSE

337   (optNSE), the correlation between the values obtained for NSE and KGE is lower (Fig. 7c). The

338   reasons for this will become much clearer later in this section, but briefly it is useful to keep in mind

339   that optimization on KGE strongly controls the values that the α and β components can achieve,

340   whereas optimization on NSE constrains these components only weakly.

341   **Fig. 7 near here**

342   The relative contributions of the criterion components to the overall model performance obtained

343   via optimization are shown in Fig. 8 (see Eq. 6 for optNSE and Eq. 11 for optKGE). The obtained

344   (optimized) model performance is dominated by the component representing r (dark grey), whereas

345   the other components representing the bias (light grey) and the variability (medium grey) of flows

346   have only small relative contributions. This applies for all 49 basins and for both optimization on

347   NSE (Fig. 8a) and KGE (Fig. 8b). However, a low relative contribution of a component to the final

348   value of the (optimized) model performance does not necessarily imply that the model performance

349   criterion is, in general, insensitive to this component. Instead, the relative contribution of a

350   component can be small because of (1) low 'weight' of the component in the equation for

351   calculating the overall model performance, and/or (2) the value of the component is close to its

352   optimal value. As a consequence of (2), the relative contribution of the components representing the

353   bias and the variability of flows can become large for non-optimal parameter sets.

354   To illustrate these considerations, Fig. 8c and Fig. 8d show the relative contribution of the criterion

355   components using random parameter sampling for a selected basin (Glan River). The sampled

356   points are arranged from left to right in order of decreasing performance for the selected criterion.

357   With decreasing overall model performance (either NSE or KGE) there is a general tendency for the

358   relative contribution of r to decrease and for the other two components to become much more

359   important. In some cases only the component representing the bias is dominant, whereas in other

360   cases only the component representing the variability of flows is dominant. This clearly indicates

361   that both NSE and KGE are sensitive to all three of the components. From a multi-objective point of

362   view this is definitely desirable, because it means that by calibrating on the overall model

363   performance we can substantially improve the components representing the bias and the variability

364   of flows. Here of course we should remember that in NSE the bias is normalized by the standard

365 deviation of the observed flows and that the 'optimal' α is equal to r. Hence, with NSE it is not

366 necessarily assured that from a hydrological point of view good values for α and β are obtained.

367 **Fig. 8 near here**

368 The cumulative distribution functions for the NSE, r, α, and β measures as obtained with optNSE

369 and optKGE in the calibration and evaluation periods are shown in Fig. 9. Looking first at the

370 results for the NSE criterion (Fig. 9a), we see that while the NSE obtained by optNSE is larger than

371 with optKGE, the difference is rather small. This indicates that by calibrating on KGE, we have

372 obtained only a slight deterioration in overall performance as measured by NSE. Further, although

373 there is a pronounced reduction in NSE from calibration to evaluation period, the reduction is

374 similar for both optNSE and optKGE.

375 However, the change in NSE tells us little that is diagnostically useful about the causes of this

376 'deterioration' in overall model performance. Of more interest, are the values obtained for the three

377 criterion components. The results for the calibration period are discussed first. Note that the

378 distribution of r is almost identical with either optNSE or optKGE (Fig. 9b, filled symbols),

379 indicating that both of the criteria have achieved similar hydrograph match in terms of shape and

380 timing. However, for the other two components, optKGE has achieved considerably better results.

381 Fig. 9c shows that there is a strong tendency for underestimation of α by optNSE (filled circle

382 symbols), due to which only 18 percent of the basins are within 10 percent of the ideal value at

383 unity, whereas for optKGE (filled triangle symbols) all of the basins are within 10 percent of the

384 ideal value. Similarly optKGE yields good results for β (Fig. 9d), with all of the basins having a

385 bias of less than 10 percent, while for optNSE 16 percent of the basins have a bias of greater than

386 10 percent. In general, optKGE results in a β value that is much closer to the ideal value at unity

387 than with optNSE. Thus, the use of optKGE has resulted in all of the basins having α and β close to

388   their ideal values of unity during calibration. This now explains why we get such a high correlation

389   between NSE and KGE in Fig. 7d; because both α and β are now almost constant across the basins

390   (here close to unity), the equations for KGE and NSE both become approximately linear functions

391   of r, and in fact we tend towards the relationship $\text{NSE}(\theta_{optKGE}) = 2*\text{KGE}(\theta_{optKGE})-1$.

392   Next we examine what happens for the evaluation period. In general, we see that the statistical

393   distributions of the three components have changed. The cumulative distribution function of r has

394   shifted to lower values in a consistent manner for both optNSE and optKGE (Fig. 9b), so that both

395   methods yield again very similar results for timing and shape. However, the optKGE calibrations

396   have retained a median value of α close to unity (the same as during calibration) while the overall

397   variability in the distribution has increased around the median value (Fig. 9c). This indicates that

398   the statistical tendency to provide good reproduction of flow variability *persists* into the evaluation

399   period, but there is an increase in the noise so that the distribution has become much wider. In

400   contrast, the optNSE results continue to show a systematic tendency to underestimate α (variability

401   of flows) during the evaluation period along with a considerable increase in random noise.

402   Similarly, the cumulative distribution function of β obtained by both methods remains centred close

403   to its calibration value while showing an increase in the variability (Fig. 9d). The small shift in the

404   median value may be caused by the fact that there is approximately 5 % less precipitation during the

405   evaluation period. Clearly, the KGE criterion has provided model calibrations that are statistically

406   more desirable during calibration while providing results that remain statistically more consistent on

407   the independent evaluation period.

408   **Fig. 9 near here**

409   An interesting observation is that in a few basins the paradoxical case occurs where *all three of the*

410   *criterion components improve with optKGE, but the value of NSE decreases* when compared to the

411   NSE obtained with optNSE (Table 2). The reason for this is the interplay between the terms α and r

412 in the NSE equation (illustrated nicely in Fig. 1). It is therefore actually (counter intuitively)

413 possible for both α and r to get closer to unity while NSE gets smaller. This is, of course, because

414 optimization on NSE seeks to make $\alpha = r$, and therefore '*punishes*' solutions for which α is close to

415 the ideal value of unity, while r will always be smaller than unity.

416 **Table 2 near here**

417 As discussed earlier, it is likely that optimization with NSE will yield results where α is close to r.

418 Fig. 10a shows a comparison between r and α obtained by the two optimization cases for all of the

419 basins. In general, when optimizing with NSE, the value of α is indeed very similar to r, which

420 means that the variability of flows is systematically underestimated (as shown above), and α

421 approaches the ideal value of unity in only one of the 49 basins. In contrast, when optimizing with

422 KGE, the value of α is close to the ideal value of unity for most of the basins.

423 Consequently, as expected from the theoretical discussion, systematically different results are

424 obtained by optNSE and optKGE for the slopes of the regression lines (Fig. 10b), where the cases

425 of regressing the simulated against the observed values (k_o, Eq. 14) and regressing the observed

426 against the simulated values (k_s, Eq. 13) are distinguished. In general, when using optNSE the value

427 of k_s is close to the ideal value at unity, but k_o is significantly smaller than one. In the case of

428 optKGE both k_s and k_o are smaller than one, but the underestimation is not as large as for k_o with

429 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

430 to be equal to unity, which would only happen if the model and data were perfect.

431 **Fig. 10 near here**

432 Finally, we report briefly on the optimal parameter values obtained using optNSE and optKGE.

433 Interestingly, even though the statistical properties of the streamflow hydrographs (as measured by

434 α and β) did change significantly (Fig. 9), for many basins the *parameter values* did not change by

435   large amounts (compared to the feasible parameter range) when moving from optNSE to optKGE

436   (Fig. 11). The correlation between the parameter values of optNSE and optKGE is at least 0.80 for

437   all six of the parameters, and for three of the parameters the correlation is larger than 0.90. For the

438   parameter $K1$ the values are slightly smaller with optKGE, which has the effect of higher peaks and

439   quicker recession of surface flow. Also the parameter $K3$ decreases with optKGE, which has the

440   effect of a less dampened base flow response. Given the function of these two parameters in the

441   model structure, a reduction in the parameter values has the effect of increasing the value of the α

442   measure. In addition, we see an increase in the percolation parameter $K2$, which results in more

443   surface flow and less base flow, with the overall effect of increasing the value of α.

444   The function of the parameters $S1_{max}$ and $Beta$ in the model is mainly to control the partitioning of

445   precipitation into runoff and evapotranspiration (thereby controlling the water balance), and as a

446   consequence these parameters mainly affect the β measure. However, these parameters also affect

447   the α measure and parameter interaction between $S1_{max}$ and $Beta$ complicates the analysis. Given

448   the function of these parameters in the model, the β measure should increase with a decrease in

449   either $S1_{max}$ and/or $Beta$, but this is not obvious from Fig. 11, because a decrease in $S1_{max}$ can be

450   compensated by an increase in $Beta$, and vice versa.

451   For the parameter $S2_{crit}$ no clear tendency of change is visible. Here it should be mentioned that

452   there was no change in the parameter values in sixteen of the basins for which the parameter values

453   were at their lower bounds (4 basins) and upper bounds (12 basins), respectively. Note that these 16

454   points also contribute to the rather high correlation observed.

455   **Fig. 11 near here**

456   On a visual, albeit subjective, basis a comparison of the parameter sets obtained by optNSE and

457   optKGE reveals that in many of the basins the two parameter sets are almost indistinguishable, but

458   nevertheless the criterion components have changed. As an example, Fig. 12a displays a

459   comparison of the parameters obtained by optNSE and optKGE for the Glan River. Apparently the

460   parameter values are quite similar, although the α measure and to a lesser extent the β measure have

461   both improved when using optKGE (see Table 2). For many basins, the difference in each of the

462   parameters was found to be only a small percentage of the overall feasible range (Fig. 12b); in 14 of

463   the 49 basins, all of the six parameters have changed by less than 10%, and in only a few of the

464   basins did two or more parameters change by a significant amount. For the latter, the changes may

465   also (at least in part) be a consequence of parameter interactions; for example, there is a clear

466   tendency for $K2$ and $S2_{crit}$ to increase/decrease simultaneously, and this fact must, of course, also be

467   considered when interpreting the scatter plots in Fig. 11.

468   **Fig. 12 near here**

469   # 4   Discussion

470   A decomposition of the NSE criterion shows that this measure of overall model performance can be

471   represented in terms of three components, which measure the linear correlation, the bias and the

472   variability of flow. By simple theoretical considerations, we can show that problems can arise in

473   model calibrations that seek to optimize the value of NSE (or its related MSE). First, because the

474   bias is normalized by the standard deviation of the observed flows, the relative importance of the

475   bias term will vary across basins (and also across years), and for cases where the variability in the

476   observed flows is high, the bias will have a low 'weight' in the computation of NSE. Second, there

477   will be a tendency for the variability in the flows to be systematically underestimated, so that the

478   ratio of the simulated and observed standard deviations of flows will tend to be equal to the

479   correlation coefficient. As a consequence, the slope of the regression line (when regressing

480   simulated against observed values) will be smaller than one, so that runoff peaks will tend to be

481   systematically underestimated. This finding may seem to contradict the general notion that

482   optimization on NSE will improve simulation of runoff peaks. In fact NSE is generally found to be

483   highly sensitive to the large runoff values, because of the (typically) larger model and data *errors*

484   involved in the matching of such events, and this fact is separate from the general (theoretical and

485   practical) tendency to underestimate the runoff peaks. Of course, when it is of interest to regress the

486   observed against the simulated values then an optimization on NSE can yield desirable results, since

487   in such a case the optimal slope of the regression line for maximizing NSE is equal to unity.

488   These theoretical considerations were all supported by the results of the modelling experiment. Of

489   course, in such an experiment, not all solutions within the theoretical criteria space are possible

490   because of constraints regarding the model structure, parameter ranges, and available data.

491   However, it was found that the simple model was capable of achieving good solutions for the bias

492   and the variability of flows with only slight decreases in the correlation coefficient. The

493   optimization task therefore becomes one of specifying the objective function in such a way that it is

494   capable of achieving such a solution as an optimal solution (i.e. simultaneously good solutions for

495   bias, flow variability and correlation). Apparently, this was not the case with NSE, and we therefore

496   formulated an alternative criterion (KGE) that is based on an equal weighting of the three

497   components (correlation, bias, and variability measures). Of course the correlation will not, in

498   general, reach its ideal value of unity, but an optimization on KGE resulted in the other two

499   components being indeed close to their ideal values. Thus, the use of KGE instead of NSE for

500   model calibration improved the bias and the variability measure considerably while only slightly

501   decreasing the correlation.

502   The simulation results were also examined for an independent evaluation period. In general, the

503   overall model performance and the individual components deteriorated in a statistical sense. It is at

504   least partially likely that this is due to the rather short lengths of the calibration and evaluation

505   periods used in this study (five and three years, respectively). Further, it should be noted that this

506   study has not accounted for either the uncertainty in the parameter values or the uncertainty in the

507   computed statistics, which would require a more rigorous Bayesian approach. Nevertheless, the

508   results clearly show that optimization using NSE tends to underestimate the variability of flows on

509   the calibration period, and that this behaviour tends to persist into the evaluation period. Further, the

510   bias in the calibration period is well constrained with KGE, but not with NSE, whereas in the

511   evaluation period (with overall poorer bias) the results with NSE are only slightly inferior to KGE.

512   An interesting result is that for many basins the optimal parameter values changed by only small

513   amounts (relative to the feasible range) when using KGE instead of NSE. In the KGE optimization

514   there was a tendency to decrease the recession parameters of surface flow and base flow to simulate

515   a flashier hydrograph, and thereby improve the value of the variability measure. Because of

516   parameter interactions there was no clear tendency of a change in the parameters for the bias

517   measure. In general, this suggests that the values of multiple criteria can be improved by making

518   only small changes in the parameter values. This emphasizes the importance of the relative

519   sensitivity of the criterion components to changes in the parameter values. On the one hand, this is a

520   desirable effect during calibration, because we want to have measures that are actually sensitive to

521   the parameter values, thereby theoretically increasing parameter identifiability. On the other hand,

522   this raises important questions for parameter regionalization, because even a small 'error' in a

523   parameter value could result in poor values of individual measures, thereby causing poor overall

524   model performance.

525   The attempt to explain the relationships between changes in the parameters and values of the

526   criterion components relates to the idea of diagnostic model evaluation, as proposed by Gupta et al.

527   (2008) and tested by Yilmaz et al. (2008) and Herbst et al. (2009). The idea behind diagnostic

528   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 recessions 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.

552 Before entering into our concluding remarks, we should point out that the primary purpose of this

553 study was not to design an improved measure of model performance, but to show clearly that there

554 are systematic problems inherent with any optimization that is based on mean squared errors (such

555 as NSE). The alternative criterion KGE was simply used for illustration purposes. An optimization

556 on KGE is equivalent to selecting a point from the three-dimensional Pareto front with the minimal

557 distance from the ideal point. Many different alternative criteria would also be sensible, but

558 ultimately it has to be understood that each single measure of model performance has its own

559 peculiarities and trade-offs between components. In the case of KGE probably the most problematic

560 characteristic is that the slope of the regression lines will tend to be smaller than one, albeit not as

561 strongly as with NSE (when regressing simulated against observed values). Because of the simple

562 design of the KGE criterion it is straightforward to understand the trade-offs between the

563 correlation, the bias and the variability measure. These trade-offs are more complex in the case of

564 NSE.

565 If single measures of model performance are used we deem it to be imperative to clearly know the

566 limitations of the selected criterion. It then will depend upon the type of application whether these

567 limitations are of concern or not. The decomposition presented here highlights the fact that identical

568 values of the NSE criterion are not necessarily indistinguishable - as is commonly (and erroneously)

569 assumed in the literature in arguments relating to equifinality (Beven and Binley 1992, Beven and

570 Freer 2001) - since the criterion components may be quite different. Thus, when evaluating or

571 reporting results based on calibration with NSE, information about the correlation, bias, and

572 variability of flows should also be given (interestingly, this was already proposed by Legates and

573 McCabe (1999), although they did not discuss the interrelation between NSE and its three

574 components). Ultimately the decision to accept or reject a model must be made by an expert

575 hydrologist, where such a decision is best based in a multiple-criteria framework. To this end, an

576 analysis of the components that constitute the overall model performance can significantly enhance

577 our understanding of model behaviour and provide insights helpful for diagnosing differences

578 between models, basins and time periods within a hydrological context.

5 Summary and conclusions

580 In this study a decomposition of the widely used Nash-Sutcliffe efficiency (NSE) was applied to

581 analyse the different components that constitute NSE (and hence MSE). We present theoretical

582 considerations that serve to highlight problems associated with the NSE criterion. The results of a

583 case study, where a simple precipitation-runoff model was applied in several basins, support the

584 theoretical findings. For comparison we show how an alternative measure of model performance

585 (KGE) can overcome the problems associated with NSE.

586 In summary, the main conclusions of this study are:

587 - The mean squared error and its related NSE criterion consists of three components,

588   representing the correlation, the bias and a measure of variability. The decomposition

589   shows that in order to maximize NSE the variability has to be underestimated. Further, the

590   bias is scaled by the standard deviation in the observed values, which complicates a

591   comparison between basins.

592 - Given that NSE consists of three components, an alternative model performance criterion

593   KGE is easily formulated by computing the Euclidian distance of the three components

594   from the ideal point, which is equivalent to selecting a point from the three-dimensional

595   Pareto front. Such an alternative criterion avoids the problems associated with NSE (but

596   also introduces new problems).

597   •   The slopes of the regression lines are directly related with the three components. NSE is

598   suitable if the interest is in regressing the observed against the simulated values, but less

599   suitable for regressing the simulated against the observed values. This means that if NSE is

600   used in optimization, then runoff peaks will tend to be underestimated. The same applies for

601   KGE, but the underestimation will not be as severe.

602   •   After optimization, the component representing the linear correlation dominates the model

603   performance criterion for both NSE and KGE. For non-optimal parameters sets any of the

604   three components can be dominant in NSE or KGE.

605   •   Even with a simple precipitation-runoff model it is possible to obtain runoff simulations

606   where the mean and variability of flows are matched well, and the linear correlation is still

607   high. However, this applies only for optimization with KGE, since NSE does not consider

608   such a solution as 'good'.

609   •   The optimal parameter values may, in practice, only change by small amounts when using

610   KGE instead of NSE as the objective function for optimization (as in our example). This

611   emphasizes the importance of considering the sensitivity of the three components to

612   perturbations in the parameter values.

613   This study reinforces the argument that model calibration is a multi-objective problem (Gupta et al

614   1998), and shows that a decomposition of the calibration criterion into components can help to

615   greatly enhance our understanding of the overall model performance (and, by extension, the

616   differences in model performance between model structures, basins and time periods). To compute

617   these components is a straightforward task and should be included in any evaluation of model

618   simulations. Ultimately, such an approach may help in the design of diagnostically powerful

619   evaluation strategies that properly support the identification of hydrologically consistent models.

6 Acknowledgements

7 References

Anderton S, Latron J, Gallart F. 2002. Sensitivity analysis and multi-response, multi-criteria evaluation of a physically based distributed model. Hydrological Processes 16: 333-353

Beldring S. 2002. Multi-criteria validation of a precipitation-runoff model. Journal of Hydrology 257: 189-211

Bergström S. 1995. The HBV model. In Computer Models of Watershed Hydrology, Singh VP (ed.), Water Resources Publications, Highlands Ranch, CO, USA, ISBN No. 0-918334-91-8

Bergström S, Graham LP. 1998. On the scale problem in hydrological modelling. Journal of Hydrology 211: 253-265

Bergström S, Lindström G, Pettersson A. 2002. Multi-variable parameter estimation to increase confidence in hydrological modelling. Hydrological Processes 16: 413-421

Beven K, Binley A. 1992. The future of distributed models: Model calibration and uncertainty prediction. Hydrological Processes 6: 279-298

641 Beven KJ, Freer J. 2001. Equifinality, data assimilation, and uncertainty estimation in mechanistic

642 modelling of complex environmental systems. Journal of Hydrology 249: 11–29

643 BMLFUW. 2007. Hydrological Atlas of Austria, 3rd edition. Wien: Bundesministerium für Land-

644 und Forstwirtschaft, Umwelt und Wasserwirtschaft, ISBN 3-85437-250-7

645 Boyle DP. 2000. Multicriteria calibration of hydrological models. PhD Dissertation, Department of

646 Hydrology and Water Resources, The University of Arizona, Tucson, USA

647 Boyle DP, Gupta HV, Sorooshian S. 2000. Toward improved calibration of hydrologic models:

648 Combining the strengths of manual and automatic methods. Water Resources Research 36(12):

649 3663-3674

650 Cao W, Bowden WB, Davie T, Fenemor A. 2006. Multi-variable and multi-site calibration and

651 validation of SWAT in a large mountainous catchment with high spatial variability. Hydrological

652 Processes 20: 1057-1073

653 Criss RE, Winston WE. 2008. Do Nash values have value? Discussion and alternate proposals.

654 Hydrological Processes 22: 2723-2725

655 Duan Q, Sorooshian S, Gupta V. 1992. Effective and efficient global optimization for conceptual

656 rainfall-runoff models. Water Resources Research 28/4: 1015-1031

657 Garrick M, Cunnane C, Nash JE. 1978. A criterion of efficiency for rainfall-runoff models. Journal

658 of Hydrology 36: 375-381

659 Gupta HV, Sorooshian S, Yapo PO. 1998. Toward improved calibration of hydrologic models:

660 Multiple and noncommensurable measures of information. Water Resources Research 34/4: 751-

661 763

662   Gupta HV, Wagener T, Liu Yuqiong. 2008. Reconciling theory with observations: elements of a

663   diagnostic approach to model evaluation. Hydrological Processes 22: 3802-3813

664   Herbst M, Gupta HV, Casper MC. 2009. Mapping model behaviour using Self-Organizing Maps.

665   Hydrology and Earth System Sciences 13: 395-409

666   Hock R. 2003. Temperature index melt modelling in mountain areas. Journal of Hydrology 282:

667   104-115

668   Kling H, Fürst J, Nachtnebel HP. 2007. Seasonal water balance. In Hydrological Atlas of Austria,

669   BMLFUW (ed.), 3rd edition, map sheet 7.2, Wien: Bundesministerium für Land- und

670   Forstwirtschaft, Umwelt und Wasserwirtschaft, ISBN 3-85437-250-7

671   Kling H, Gupta HV. 2009. On the development of regionalization relationships for lumped

672   watershed models: The impact of ignoring sub-basin scale variability. Journal of Hydrology

673   (submitted)

674   Krause P, Boyle DP, Bäse F. 2005. Comparison of different efficiency criteria for hydrological

675   model assessment. Advances in Geosciences 5: 89-97

676   Legates DR, McCabe GJ. 1999. Evaluating the use of "goodness-of-fit" measures in hydrologic and

677   hydroclimatic model evaluation. Water Resources Research 35: 233-241

678   Lindström G. 1997. A simple automatic calibration routine for the HBV model. Nordic Hydrology

679   28: 153-168

680   Madsen H. 2003. Parameter estimation in distributed hydrological catchment modelling using

681   automatic calibration with multiple objectives. Advances in Water Resources 26: 205-216

682 Marce R, Ruiz CE, Armengol J. 2008. Using spatially distributed parameters and multi-response

683 objective functions to solve parameterization of complex applications of semi-distributed

684 hydrological models. Water Resources Research 44: 18pp

685 Martinec J, Rango A. 1989. Merits of statistical criteria for the performance of hydrological models.

686 Water Resources Bulletin, AWRA 25(2): 421-432

687 Mathevet T, Michel C, Andreassian V, Perrin C. 2006. A bounded version of the Nash-Sutcliffe

688 criterion for better model assessment on large sets of basins. In Large Sample Basin Experiment for

689 Hydrological Model Parameterization: Results of the Model Parameter Experiment - MOPEX,

690 Andréassian V, Hall A, Chahinian N, Schaake J (eds.), IAHS Publ. 397

691 McCuen RH, Snyder WM. 1975. A proposed index for comparing hydrographs. Water Resources

692 Research 11(6): 1021-1024

693 McCuen RH, Knight Z, Cutter AG. 2006. Evaluation of the Nash-Sutcliffe efficiency index. Journal

694 of Hydrologic Engineering 11(6): 597-602

695 Merz R, Blöschl G. 2004. Regionalisation of catchment model parameters. Journal of Hydrology

696 287: 95-123

697 Murphy A. 1988. Skill scores based on the mean square error and their relationships to the

698 correlation coefficient. Monthly Weather Review 116: 2417-2424

699 Nash JE, Sutcliffe JV. 1970. River flow forecasting through conceptual models. Part I. A discussion

700 of principles. Journal of Hydrology 10: 282-290

701 Parajka J, Merz R, Blöschl G. 2005. A comparison of regionalisation methods for catchment model

702 parameters. Hydrology and Earth System Sciences 9: 157-171

703   Rode M, Suhr U, Wriedt G. 2007. Multi-objective calibration of a river water quality model -

704   Information content of calibration data. Ecological Modelling 204: 129-142

705   Rojanschi V, Wolf J, Barthel R, Braun J. 2005. Using multi-objective optimisation to integrate

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

707   Schaefli B, Gupta HV. 2007. Do Nash values have value? Hydrological Processes 21: 2075-2080

708   Thornthwaite CW, Mather JR. 1957. Instructions and tables for computing potential evaporation

709   and the water balance. Publications in Climatology 10(3): 311pp

710   Tolson BA, Shoemaker CA. 2007. Dynamically dimensioned search algorithm for computationally

711   efficient watershed model calibration. Water Resources Research 43: 16 pp

712   van Griensven A, Bauwens W. 2003. Multiobjective autocalibration for semidistributed water

713   quality models. Water Resources Research 39(12): 9pp

714   Wang G, Xia J, Chen J. 2009. Quantification of effects of climate variations and human activities

715   on runoff by a monthly water balance model: A case study of the Chaobai River basin in northern

716   China. Water Resources Research 45: 12pp

717   Weglarczyk S. 1998. The interdependence and applicability of some statistical quality measures for

718   hydrological models. Journal of Hydrology 206: 98-103

719   Yilmaz KK, Gupta HV, Wagener T. 2008. A process-based diagnostic approach to model

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

721   18pp

722   Young AR. 2006. Stream flow simulation within UK ungauged catchments using a daily rainfall-

723   runoff model. Journal of Hydrology 320: 155-172

724

725

726 Table 1: Parameters of the model. Parameters in brackets were not calibrated.

parameter	units	feasible range	description
$S1_{max}$	mm	50 - 700	soil storage capacity
Beta	/	0.1 - 25	exponent for computing runoff generation
($S1_{crit}$)	/	(0.6)	critical soil moisture for actual evapotranspiration
K1	h	10 - 500	recession coefficient for surface flow
K2	h	10 - 1000	recession coefficient for percolation
$S2_{crit}$	mm	0 - 15	outlet height for surface flow
K3	h	500 - 10000	recession coefficient for base flow
(K4)	h	(0 - 10)	recession coefficient for distributed routing

727

728 Table 2: 'Paradoxical' examples for NSE and components in three basins (results for the calibration
729 period). All three components (r, α, β) improve but the overall model performance measured by
730 NSE decreases with the parameter set obtained by optKGE.

basin	method	NSE [/]	KGE [/]	r [/]	α [/]	β [/]
Zaya River	optNSE	0.484	0.685	0.714	0.871	1.019
	optKGE	0.452	0.732	0.733	1.026	1.001
Pitten River	optNSE	0.742	0.828	0.863	0.899	1.028
	optKGE	0.730	0.865	0.866	1.004	1.016
Glan River	optNSE	0.786	0.855	0.888	0.912	1.028
	optKGE	0.776	0.888	0.889	1.002	1.007

731

732

733   Figure captions:

734   Fig. 1: Relationship of NSE with α and r (β_n is assumed to be zero): (a) theoretical relationship, (b)

735   illustrative example obtained by random parameter sampling with a hydrological model (Leaf

736   River, Mississippi, USA, 1924 km^2, 11 years daily data, HyMod model; only those points where

737    $\beta_n^2 < 0.01$ are displayed). Contour lines indicate values for NSE. See colour version of this figure

738   online.

739   Fig. 2: Example for three-dimensional Pareto front of r, α and β. ED is the Euclidian distance

740   between the 'optimal' point and the ideal point, where all three measures are 1.0. Glan River,

741   Austria, 432 km^2, 5 years daily data, HBV model variant, random parameter sampling.

742   Fig. 3: Typical scatter plots depicting simulated and observed runoff ($r = 0.86$ and $\alpha = 0.90$) and

743   fitted regression lines: (a) regression against simulated runoff ($k_s = 0.96$) and (b) regression against

744   observed runoff ($k_o = 0.77$). Pitten River, Austria, 277 km^2, 5 years daily data, HBV model variant,

745   parameters optimized on NSE. Note, that in (a) and (b) the identical data points are plotted, but the

746   axes are flipped.

747   Fig. 4: Map showing locations of the 49 Austrian basins used in this study. Also depicted are the 49

748   gauges and 222 precipitation stations.

749   Fig. 5: Relationship between index of evaporation and index of wetness for the 49 Austrian basins.

750   The index of wetness is computed as the ratio between precipitation (P) and potential

751   evapotranspiration (ETp). The index of evaporation is computed as the ratio between actual

752   evapotranspiration (ETa) and ETp. Data represent long-term means from the period 1961 to 1990

753   and are taken from Hydrological Atlas of Austria (BMLFUW 2007).

754   Fig. 6: Conceptual model structure (the snow module is not shown). Parameters in brackets are not

755   calibrated.

756   Fig. 7: Scatter plots of overall model performance: cal = calibration period, eval = evaluation

757   period. Note that in (a) two points are located outside the plotting range because of negative NSE

758   values in the evaluation period.

759   Fig. 8: Stacked area plots showing the relative contribution of the components for NSE and KGE in

760   the calibration period: (a) optNSE in 49 basins, (b) optKGE in 49 basins, (c) and (d) random

761   parameter sampling in the Glan River basin.

762   Fig. 9: Cumulative distribution functions for NSE, r, α, and β as obtained with optNSE and optKGE

763   in the calibration and evaluation periods.

764   Fig. 10: Relationship between (a) r and α and (b) the slope of the regression lines k_s and k_o.

765   Fig. 11: Scatter plots of optimal parameters obtained by optNSE and optKGE. Parameter values are

766   normalized by the feasible parameter range (Table 1); the parameters *Beta*, *K1*, *K2* and *K3* are log-

767   transformed before normalization.

768   Fig. 12: Comparison of the parameter sets obtained by optNSE and optKGE: (a) normalized

769   parameter values of θ_{optNSE} and θ_{optKGE} in the Glan River basin, (b) difference in the normalized

770   parameter values (computed as $\theta_{optKGE}-\theta_{optNSE}$), displayed for all basins.

Figure 1
Click here to download high resolution image

(a) Theoretical relationship

(b) Leaf River

Figure 3

(a) regression against simulated runoff

(b) regression against observed runoff
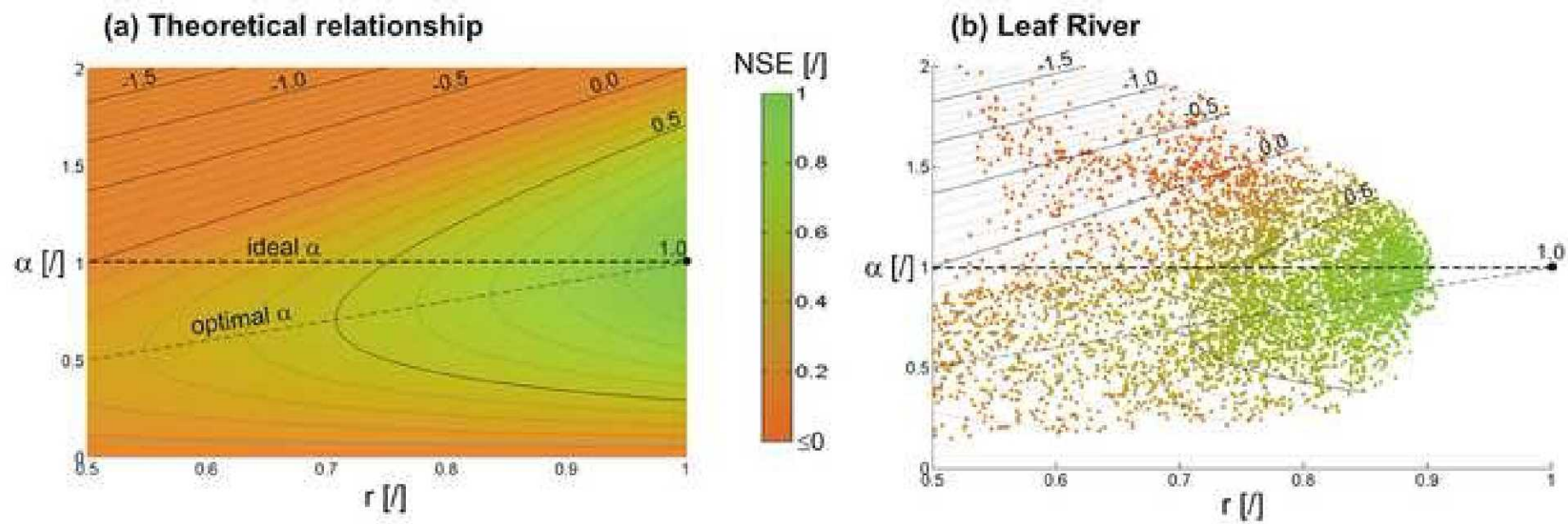
Figure 4
Click here to download high resolution image
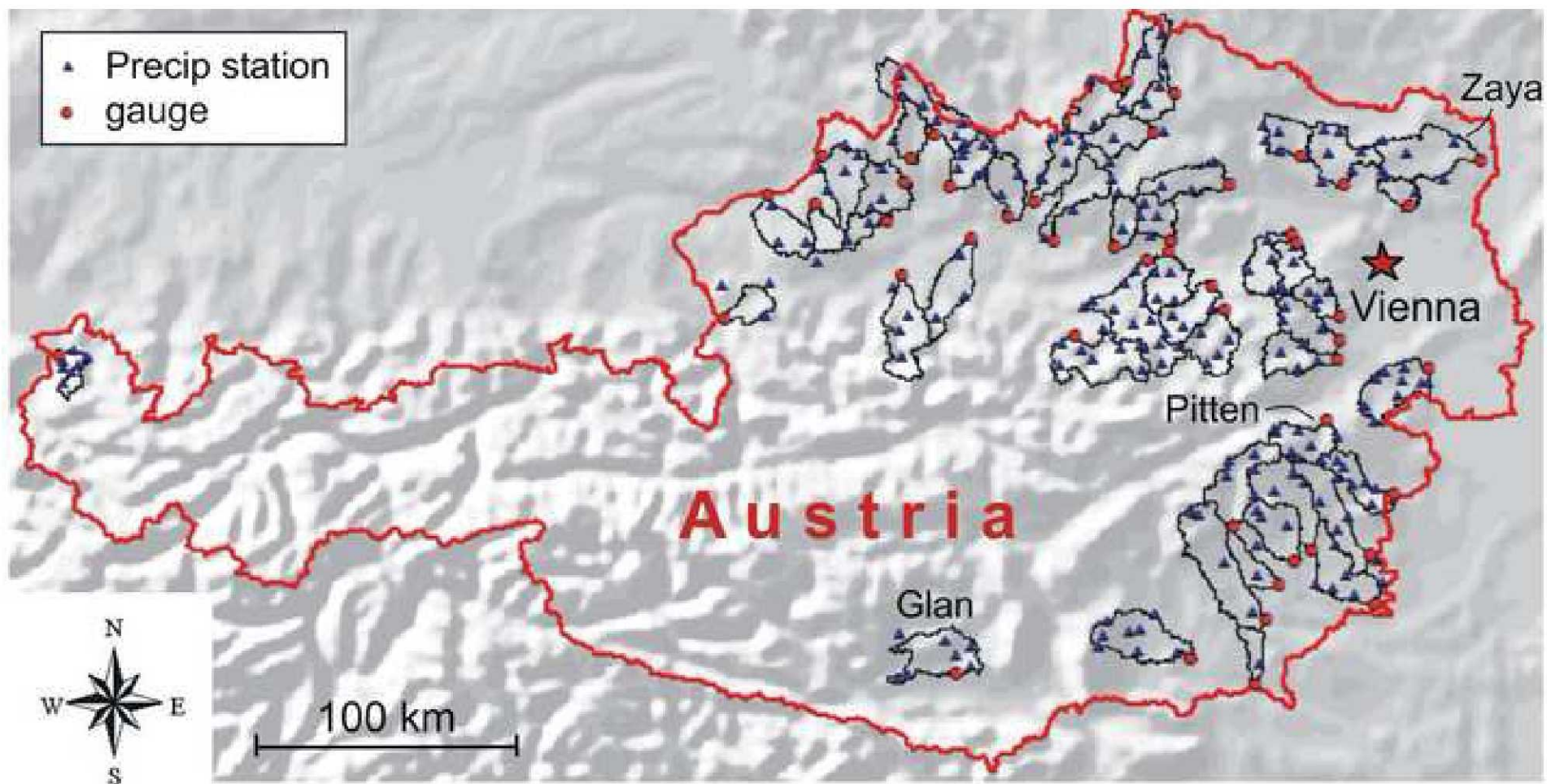
Figure 5

Figure 6
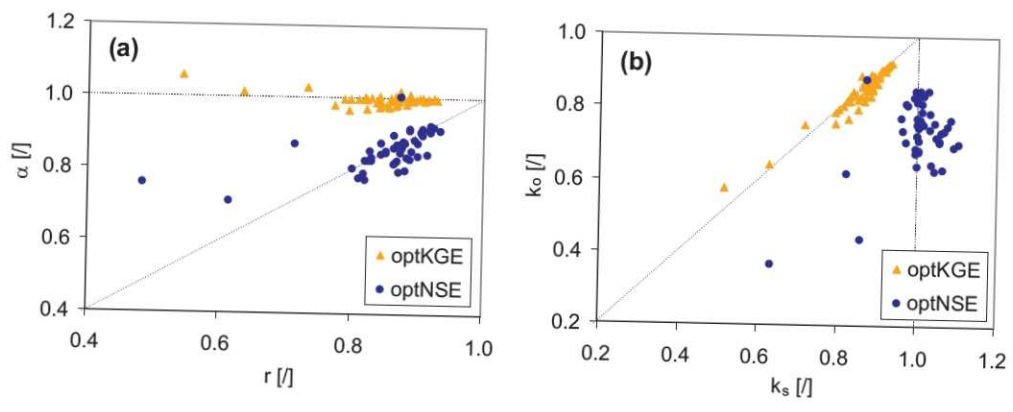
Figure 7
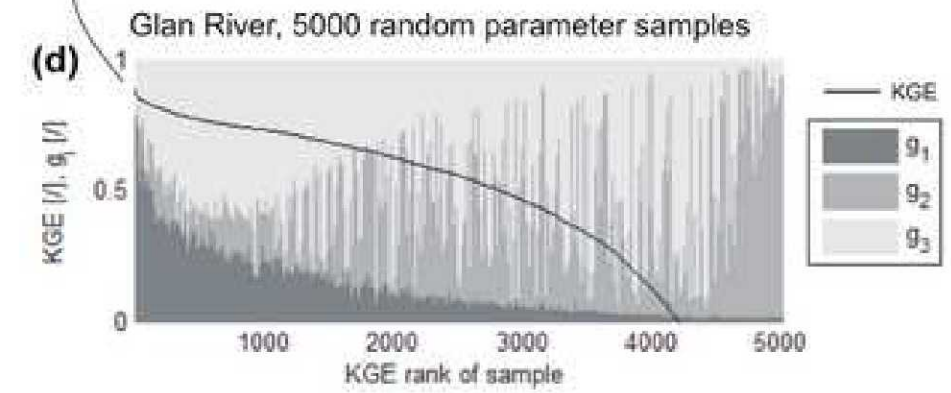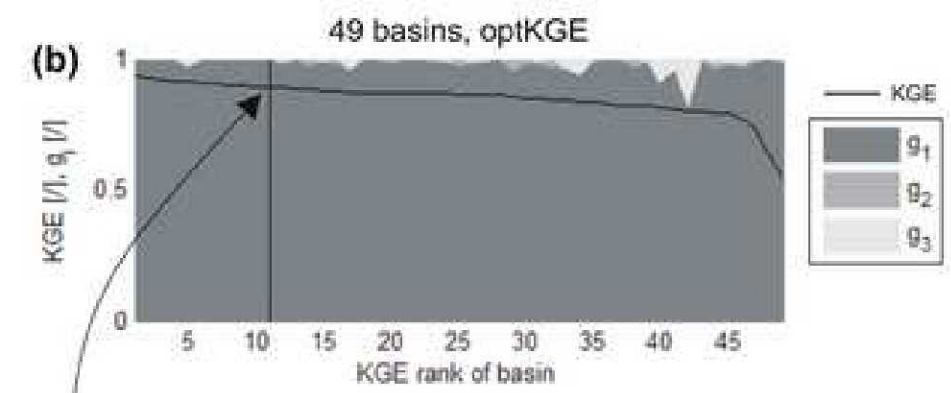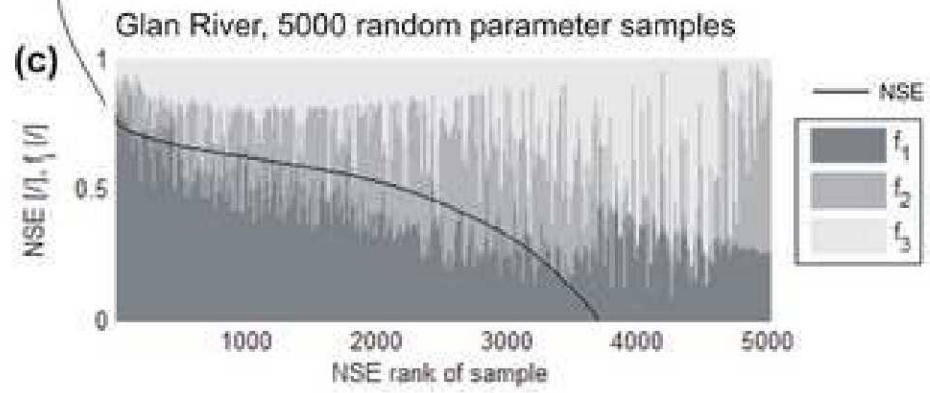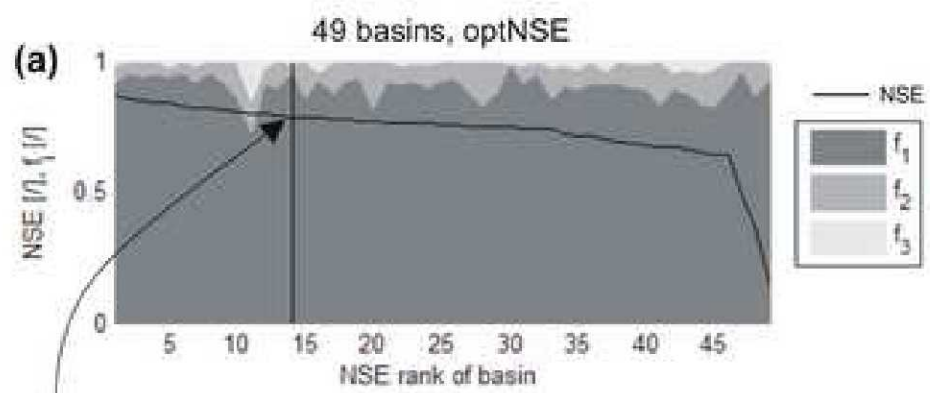
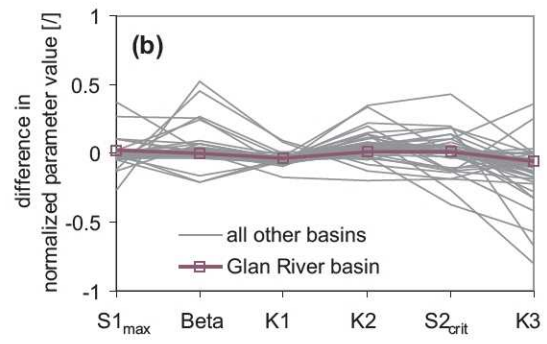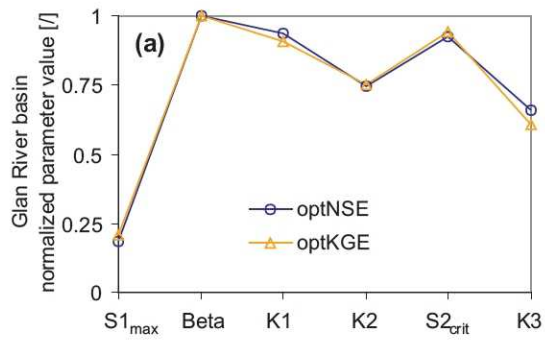**Figure 9**

Figure 10

Figure 11
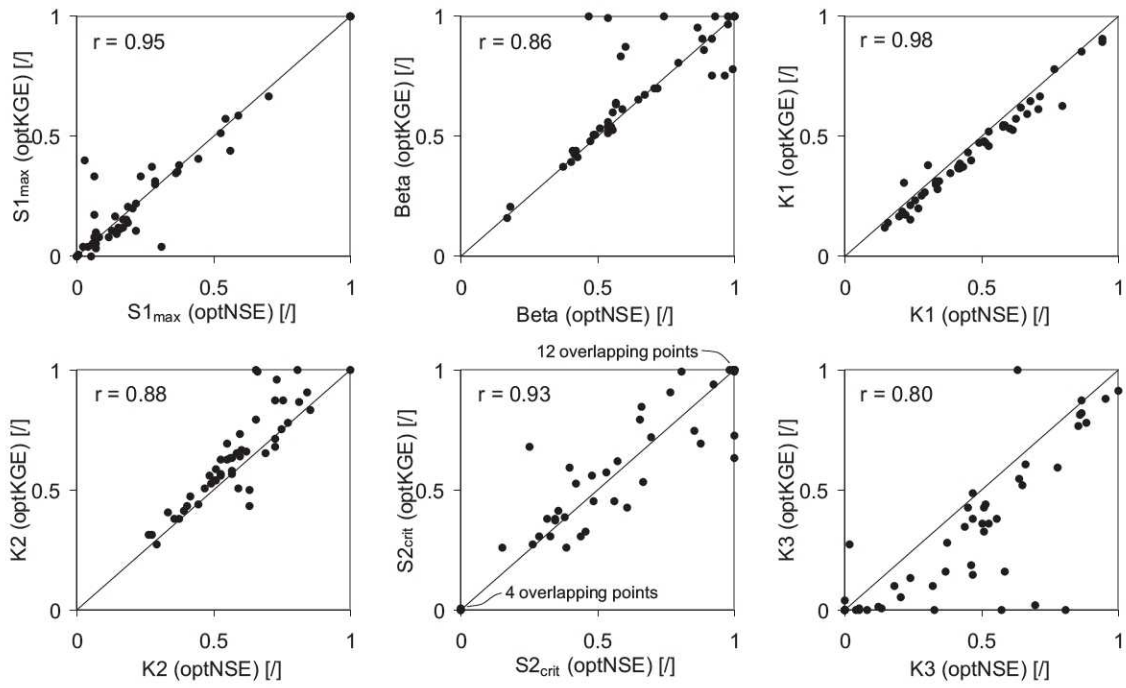
Figure 12

Gupta, H. V. Kling, H., Yilmaz, K. K., Martinez-Baquero, G.F. 2009: Decomposition of the Mean Squared Error & NSE Performance Criteria: Implications for Improving Hydrological Modelling, Journal of Hydrology.

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.