Technical note: Diagnostic efficiency – insights into model performance

Robin Schwemmle1, Dominic Demand1, Markus Weiler1

1University of Freiburg, Faculty of Environment and Natural Resources, Chair of Hydrology, Freiburg, Germany.

*Correspondence to*: Robin Schwemmle (robin.schwemmle@hydrology.uni-freiburg.de)

**Abstract.** Please use only the styles of this template (MS title, Authors, Affiliations, Correspondence, Normal for your text, and Headings 1–3). Figure 1 uses the style Caption and Fig. 1 is placed at the end of the manuscript. The same is applied to tables (Aman et al., 2014; Aman and Bman, 2015)

# 1 Introduction

Why do we need efficiency measures?

* Evaluation of model performance to quantify the prediction skill
* Model calibration

Elaborate on well-established efficiency measures (Schaefli and Gupta, 2007;Knoben et al., 2019)

KGE (Gupta et al., 2009;Kling et al., 2012;Pool et al., 2018) and NSE (Nash and Sutcliffe, 1970) return numbers between −∞ and 1, but these numbers only provide limited insights into model performance

Which studies already looked at diagnostic measures? (Yilmaz et al., 2008)

if my model performance is bad: where do the errors come from? What processes might not be captured by the model?

Diagnosing model performance by introducing a novel efficiency measure based on flow duration curve

Flow duration curve covers different processes (e.g. runoff generation, storage recession)

# 2 Methodology

Here we introduce the Diagnostic efficiency (DE, Eq. (1)). First we introduce the three components which build up the efficiency measure. The first two components are entirely based on the flow duration curve. To include the missing temporal dimension we added a third component.

The first component reflects the constant error and is represented by arithmetic mean of the relative bias (, Eq. (1)):

, (1)

i represents the exceedance probability, N the total number of data points and is the relative bias of the simulated and observed flow duration curve; = 0 indicates perfect input data; < 0 indicates underestimated input; > 0 indicates overestimated input. calculates (Eq. (2)):

, (2)

is the simulated streamflow at exceedance probability i and the observed streamflow at exceedance probability i.

The second metric component represents the dynamic error which is the absolute area of the remaining bias (, Eq. (3)):

, (3)

where the remaining bias is integrated over the entire domain of the flow duration curve. Eq. (4) is inserted in Eq. (3):

, (4)

by subtracting we remove the constant error and the dynamic error remains.

The third metric component is the Pearson correlation r for the simulated and observed discharge time series and accounts for timing errors:

(5)

Calculating the distance between , and r results in the non-normalized Diagnostic efficiency (, Eq. (6)):

, (6)

Since the mean flow benchmark of DE (; where is simulated discharge at time t) strongly depends on (i.e. is not constant across different time series; where is observed discharge at time t), we normalize DE by its (; Eq.(7)):

, (7)

The normalization is needed to enable an inter-comparison (e.g. for comparative hydrology).

However, the calculation of DE does not allow a diagnosis. Thus, we project DE in the 2-D space. First, we calculate the direction of the dynamic error (; Eq.(8)):

, (8)

where the integral of includes values from 0th percentile to 50th percentile.

In order to differentiate the dynamic error type, the slope of the remaining bias (; Eq.(9)) is computed:

, (9)

Negative values indicate that there is a tendency to overestimate high flows and underestimate low flows while positive values indicate the opposite (i.e. there is a tendency to underestimate high flows and overestimate low flows).

We used the inverse tangent to derive the ratio between and in radians ( Eq.(10)):

, (10)

Conditions for which diagnosis is possible:

,

,

,

,

Conditions to differentiate between a good model and a bad model

Kling-Gupta Efficiency (KGE; Gupta et al., 2009)

, (11)

where …

, (12)

where …

Mean flow benchmark of Kling-Gupta Efficiency (; Knoben et al., 2019)

, (13)

, (14)

Nash-Sutcliffe Efficiency (NSE; Nash and Sutcliffe, 1970)

, (15)

where ..

# 3 Proof of concept

Errors caused by:

* ineffective model parameters (Wagener and Gupta, 2005)
* model structure (Clark et al., 2008;Clark et al., 2011)
* input data (Yatheendradas et al., 2008)
* uncertainties in observations (Coxon et al., 2015) to which simulations are compared to

We used an observed streamflow time series from the CAMELS data set (Newman et al., 2015). Near-natural catchment and sufficiently long temporal coverage, could be any time series. In order to mimic model errors, we systematically manipulated the observed time series.

## 3.1 Mimicking errors

Three types of errors…dynamic error…constant error…timing error

Mimicking dynamic errors:

1. Increase high flows – Decrease low flows: Multiplying the observed time series with a vector (1.5 … 0.5)
2. Decrease high flows – Increase low flows: Multiplying the observed time series with a vector (0.5 … 1.5)

Mimicking constant errorr:

1. Positive offset: Multiplying the observed time series with a constant > 1
2. Negative offset: Multiplying the observed time series with a constant < 1

Timing error due to dynamic error and/or constant error:

1. Shuffling: Randomizing the order of the observed time series

Combination of dynamic error and constant error:

1. Decrease high flows – Increase low flows and negative offset
2. Decrease high flows – Increase low flows and positive offset
3. Increase high flows – Decrease low flows and negative offset
4. Increase high flows – Decrease low flows and positive offset

Benchmark against KGE and NSE:

1. Mean flow benchmark

Combination of dynamic error, constant error and timing error:

1. Decrease high flows – Increase low flows, negative offset and shuffling
2. Decrease high flows – Increase low flows, positive offset and shuffling
3. Increase high flows – Decrease low flows, negative offset and shuffling
4. Increase high flows – Decrease low flows, positive offset and shuffling

Perfect simulation

(‘1’) Manipulated time series corresponds to observed time series

## 3.2 Real case example

## - CAMELS dataset; evaluation of three model runs with different parameter sets but same input data

# 4 Conclusions

* tool for diagnostic model evaluation
* identifying orgin of errors visualizing the three components in a 2D-space
* Comparison to KGE and NSE
* advancing model development

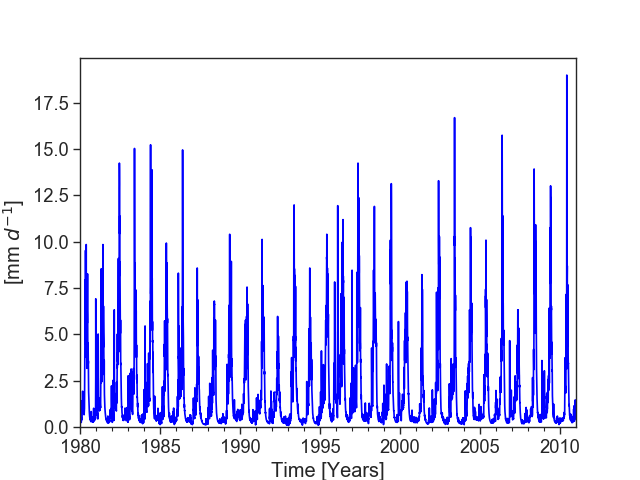


Figure 1: Observed streamflow time series from CAMELS dataset (Newman et al., 2015; gauge\_id: 13331500; gauge\_name: Minam River near Minam, OR, U.S.)

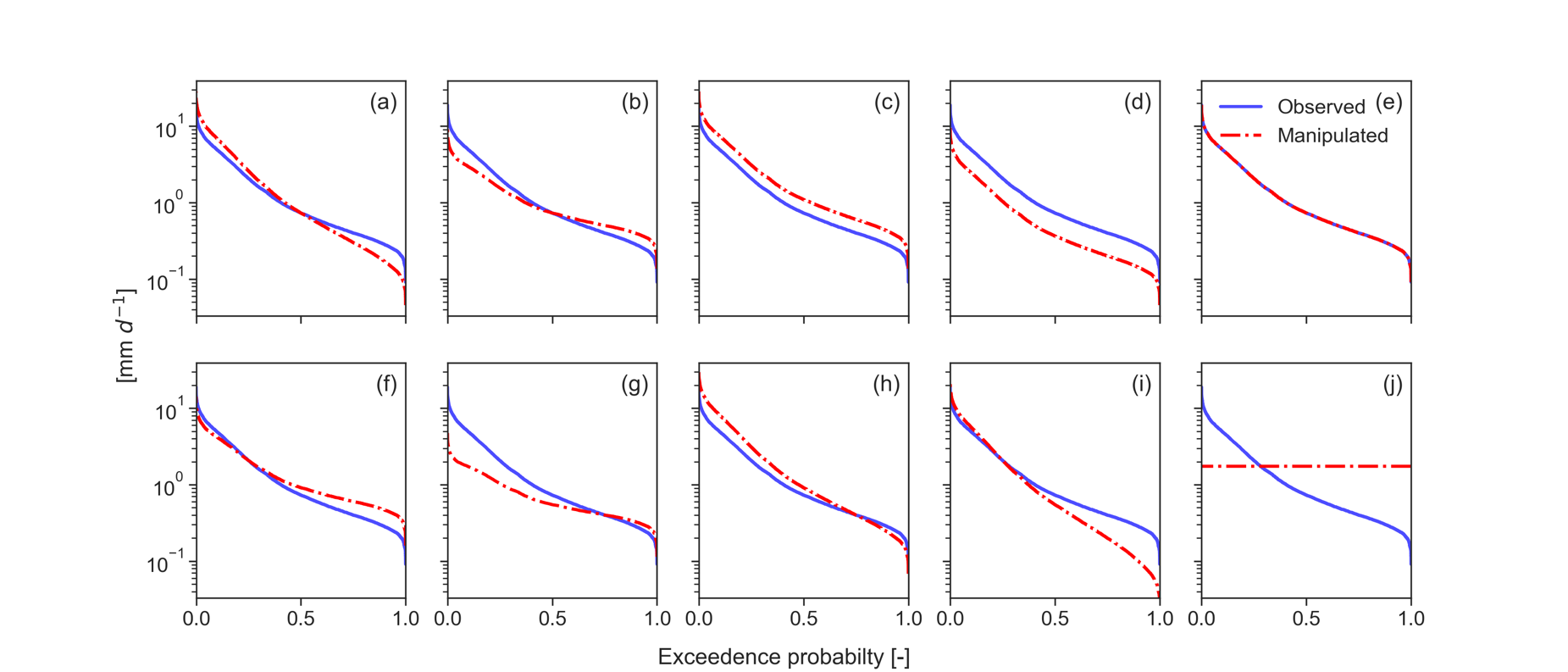


Figure 2: Flow duration curves (FDCs) of observed (blue) and manipulated (dashed red) streamflow time series. Manipulated FDCs are depicted for (a-b) dynamic errors only, (c-d) constant errors only, (e) timing error only; (f-i) combination of dynamic and constant error and (j) the mean flow benchmark. The combination of dynamic error, constant error and timing error is not shown, since they are identical to f-i.

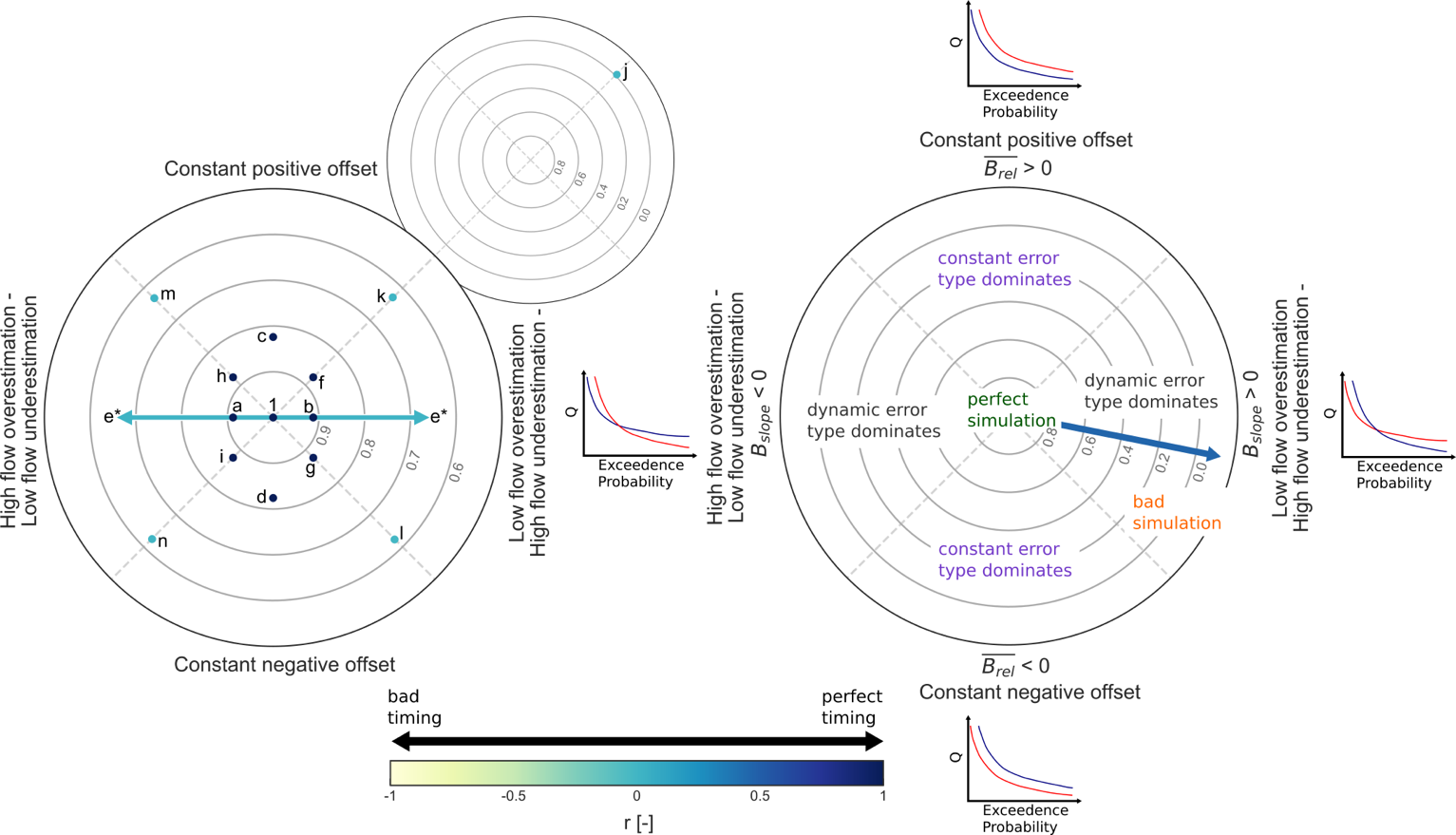
Figure 3: (left) Diagnostic polar plot for the mimicked errors (a-i; k-n) visualizing the overall model performance (*DE*; contour lines) and contribution of constant error (> 0: constant positive offset; < 0: constant negative offset), dynamic error (high flow underestimation – low flow overestimation; high flow overestimation – low flow underestimation) and timing error (*r;* blue (yellow) indicates temporal match (mismatch)). (e\*) type of dynamic error cannot be distinguished. Inset shows diagnostic polar plot for the mean flow benchmark (j). (right) Annotated diagnostic polar plot illustrating the interpretation. Hypothetic FDC plots give examples for the error types.

Table 1: Comparison of DE, KGE and NSE for mimicked errors

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | a | b | c | d | e | f | g | h | i | j | k | l | m | n | ‘1' |
| *DE* | 0.91 | 0.91 | 0.82 | 0.82 | 0.65 | 0.88 | 0.88 | 0.88 | 0.88 | 0 | 0.63 | 0.62 | 0.63 | 0.63 | 1 |
|  | 0.59 | 0.6 | 0.5 | 0.5 | 0.29 | 0.83 | 0.35 | 0.35 | 0.82 | 0 | 0.27 | 0.03 | 0.04 | 0.28 | 1 |
| *NSE* | 0.7 | 0.7 | 0.6 | 0.6 | -1 | 0.94 | 0.27 | 0.27 | 0.94 | 0 | -0.56 | -0.25 | -3.2 | -1.51 | 1 |

* Mean flow benchmark for DE is not constant
* NSE is not constant for synthetically generated errors

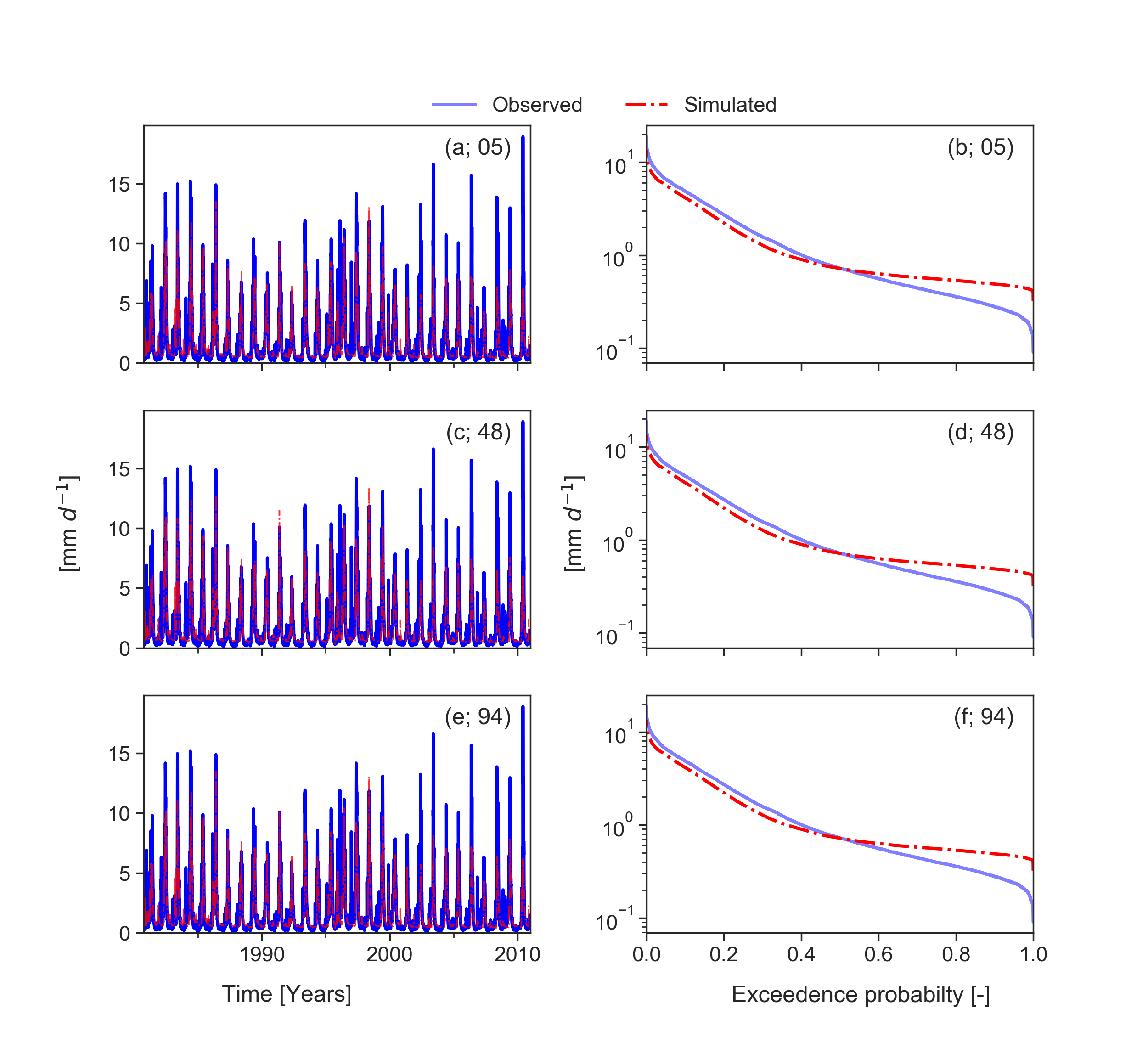


Figure 4: Simulated and observed streamflow time series of real case example (a, c and e) and the related flow duration curves (b, d and f). Time series are derived from the CAMELS dataset (Newman et al., 2015). Observed time series is the same as in Figure 1. Simulated time series had been produced by model runs with different parameter sets (set id:05, 48, 96) but same input data.

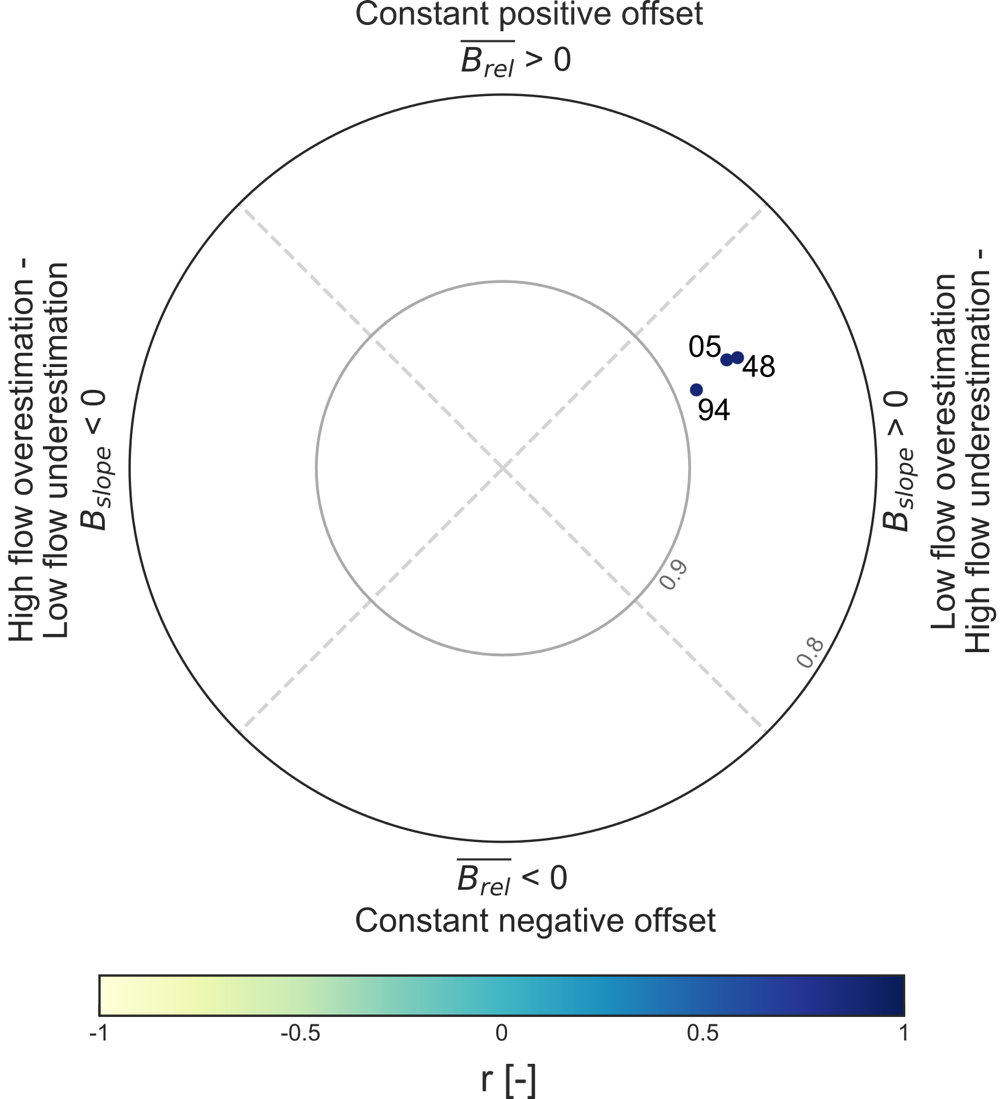


Figure 5: Diagnostic plot for real case example. Three different simulation runs are evaluated (05, 48, 94; see Figure 4). All simulations perform well. However, the remaining error is dominated by the dynamic error type while timing is excellent.

* Overall good performance, dynamic error type dominates

# References

Clark, M. P., Slater, A. G., Rupp, D. E., Woods, R. A., Vrugt, J. A., Gupta, H. V., Wagener, T., and Hay, L. E.: Framework for Understanding Structural Errors (FUSE): A modular framework to diagnose differences between hydrological models, Water Resources Research, 44, 10.1029/2007wr006735, 2008.

Clark, M. P., Kavetski, D., and Fenicia, F.: Pursuing the method of multiple working hypotheses for hydrological modeling, Water Resources Research, 47, 10.1029/2010wr009827, 2011.

Coxon, G., Freer, J., Westerberg, I. K., Wagener, T., Woods, R., and Smith, P. J.: A novel framework for discharge uncertainty quantification applied to 500 UK gauging stations, Water Resources Research, 51, 5531-5546, 10.1002/2014wr016532, 2015.

Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F.: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, Journal of Hydrology, 377, 80-91, 10.1016/j.jhydrol.2009.08.003, 2009.

Kling, H., Fuchs, M., and Paulin, M.: Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios, Journal of Hydrology, 424-425, 264-277, 10.1016/j.jhydrol.2012.01.011, 2012.

Knoben, W. J. M., Freer, J. E., and Woods, R. A.: Technical note: Inherent benchmark or not? Comparing Nash-Sutcliffe and Kling-Gupta efficiency scores, Hydrol. Earth Syst. Sci. Discuss., 2019, 1-7, 10.5194/hess-2019-327, 2019.

Nash, J. E., and Sutcliffe, J. V.: River flow forecasting through conceptual models part I - A discussion of principles, Journal of Hydrology, 10, 282-290, 10.1016/0022-1694(70)90255-6, 1970.

Newman, A. J., Clark, M. P., Sampson, K., Wood, A., Hay, L. E., Bock, A., Viger, R. J., Blodgett, D., Brekke, L., Arnold, J. R., Hopson, T., and Duan, Q.: Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: data set characteristics and assessment of regional variability in hydrologic model performance, Hydrol. Earth Syst. Sci., 19, 209-223, 10.5194/hess-19-209-2015, 2015.

Pool, S., Vis, M., and Seibert, J.: Evaluating model performance: towards a non-parametric variant of the Kling-Gupta efficiency, Hydrological Sciences Journal, 63, 1941-1953, 10.1080/02626667.2018.1552002, 2018.

Schaefli, B., and Gupta, H. V.: Do Nash values have value?, Hydrological Processes, 21, 2075-2080, 10.1002/hyp.6825, 2007.

Wagener, T., and Gupta, H. V.: Model identification for hydrological forecasting under uncertainty, Stochastic Environmental Research and Risk Assessment, 19, 378-387, 10.1007/s00477-005-0006-5, 2005.

Yatheendradas, S., Wagener, T., Gupta, H., Unkrich, C., Goodrich, D., Schaffner, M., and Stewart, A.: Understanding uncertainty in distributed flash flood forecasting for semiarid regions, Water Resources Research, 44, 10.1029/2007wr005940, 2008.

Yilmaz, K. K., Gupta, H. V., and Wagener, T.: A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model, Water Resources Research, 44, 10.1029/2007wr006716, 2008.

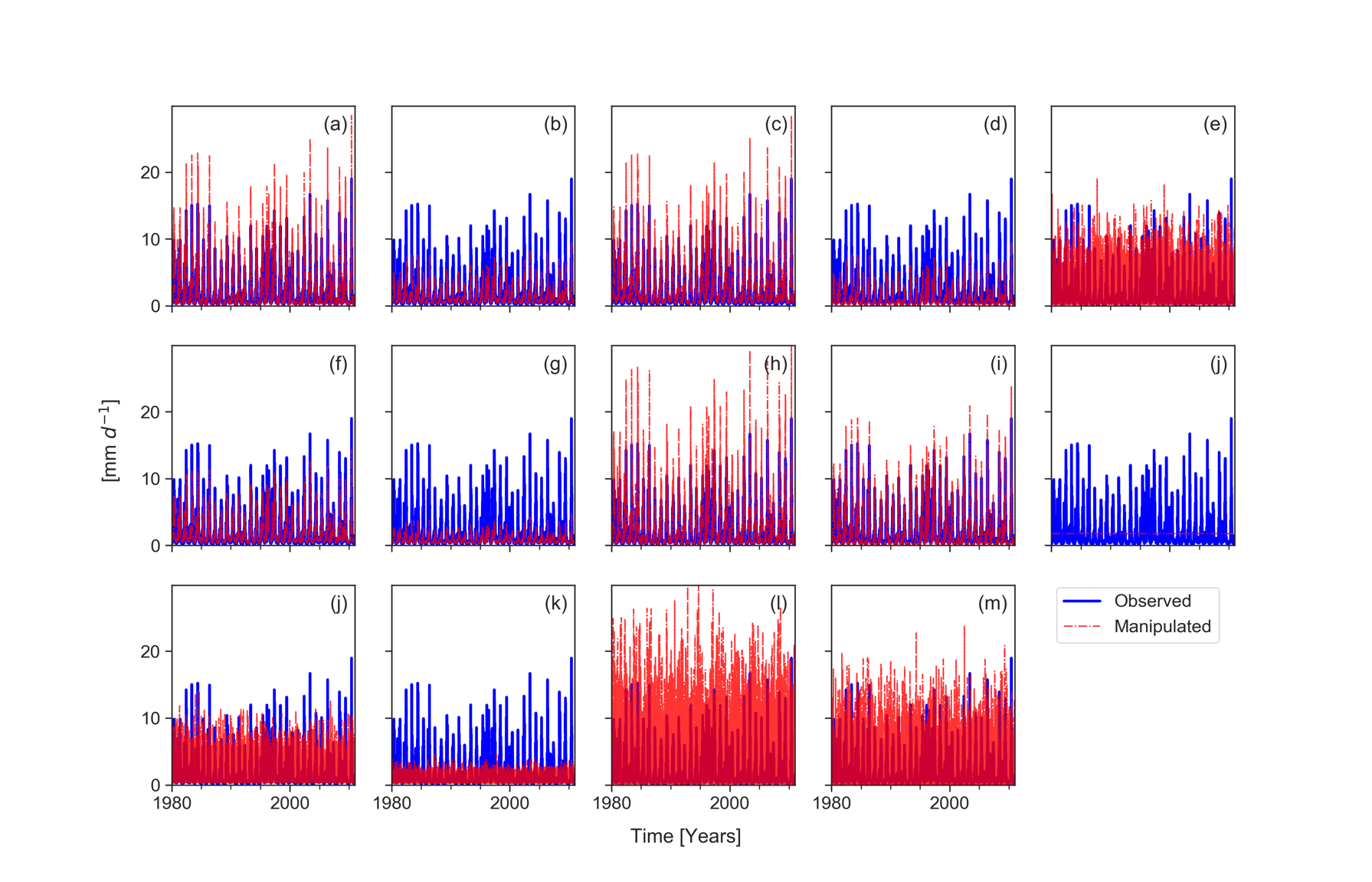


Figure A1: Time series of observed and manipulated streamflow

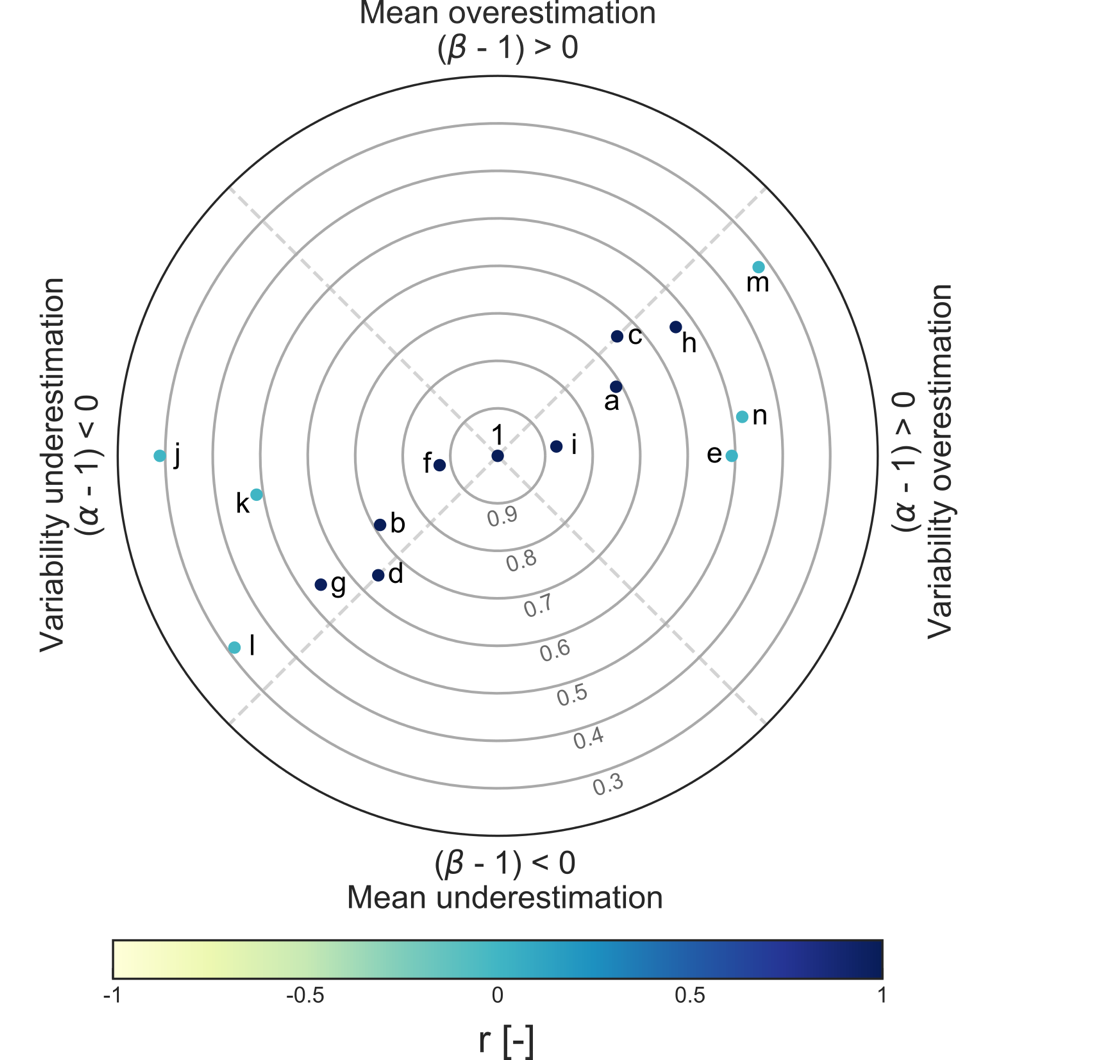


Figure A2: Polar plot of KGE

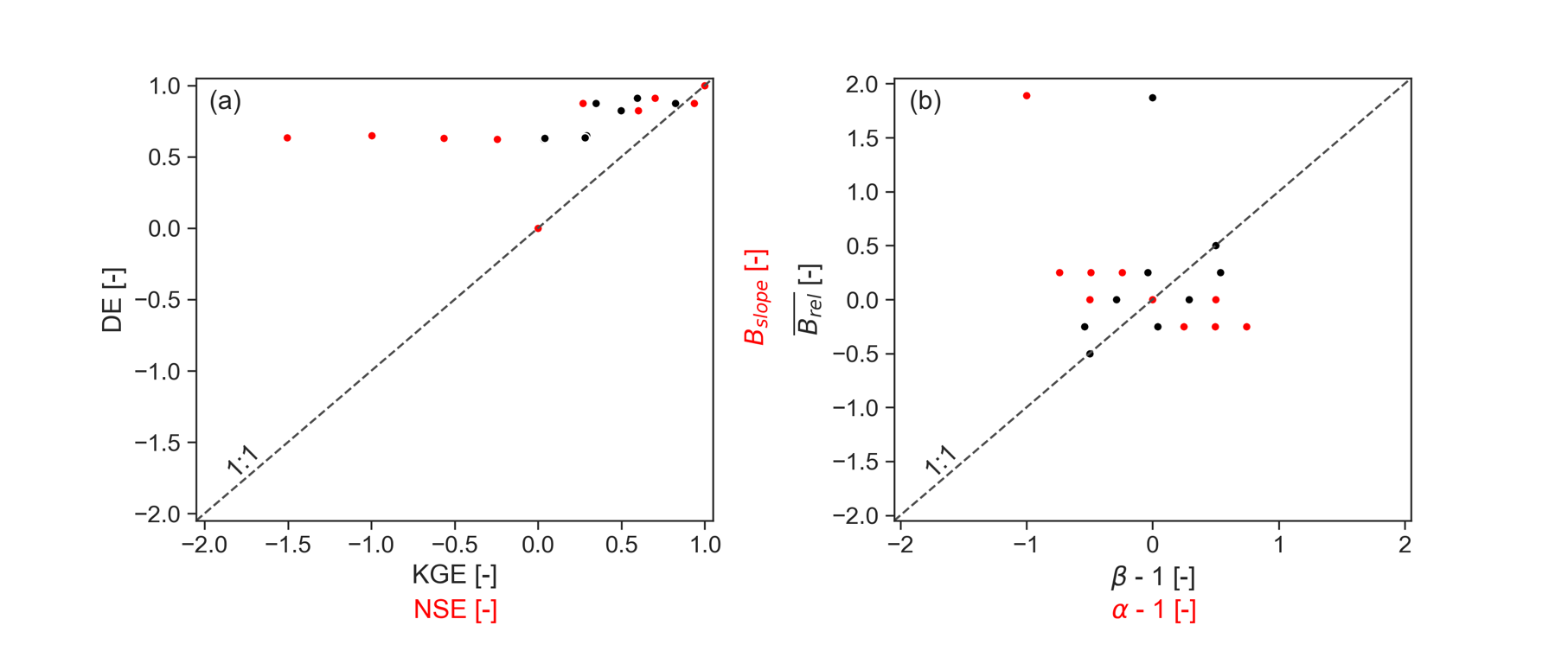


Figure A3: (a) Scatterplot to compare DE with KGE (black) and DE with NSE (red), respectively. (b) Scatterplot to compare with (black) and with (red), respectively.

Table A1: Comparison of metric components

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | a | b | c | d | e | f | g | h | i | j | k | l | m | n | 1 |
|  | 0 | 0 | 0.5 | -0.5 | 0 | 0.25 | -0.25 | 0.25 | -0.25 | 1.87 | 0.25 | -0.25 | 0.25 | -0.25 | 0 |
|  | 0.25 | 0.25 | 0 | 0 | 0 | 0.25 | 0.25 | 0.25 | 0.25 | 1.89 | 0.25 | 0.25 | 0.25 | 0.25 | 0 |
| *r* | 1 | 1 | 1 | 1 | 0 | 1 | 0.98 | 1 | 1 | 0 | 0.01 | -0.01 | 0.01 | 0.02 | 1 |
|  | 0.12 | -0.12 | 0 | 0 | 0 | -0.12 | -0.12 | 0.12 | 0.12 | -0.93 | -0.12 | -0.12 | 0.12 | 0.12 | 0 |
|  | -0.25 | 0.25 | 0 | 0 | 0 | 0.25 | 0.25 | -0.25 | -0.25 | 1.89 | 0.25 | 0.25 | -0.25 | -0.25 | 0 |
|  | 1.5 | 0.51 | 1.5 | 0.5 | 1 | 0.76 | 0.26 | 1.75 | 1.25 | 0 | 0.76 | 0.26 | 1.75 | 1.25 | 1 |
|  | 1.29 | 0.71 | 1.5 | 0.5 | 1 | 0.96 | 0.46 | 1.54 | 1.04 | 1 | 0.96 | 0.46 | 1.54 | 1.04 | 1 |

Table A2: Comparison of DE, KGE and NSE for real case example

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| set\_id |  |  | *r* | *DE* |  |  |  |  |  |  |  | *NSE* |
| 05 | 0.16 | 0.32 | 0.88 | 0.87 | -1.84 | -0.15 | 0.32 | 0.45 | 0.81 | 0.90 | 0.79 | 0.77 |
| 48 | 0.16 | 0.34 | 0.89 | 0.86 | -1.84 | -0.16 | 0.34 | 0.44 | 0.81 | 0.89 | 0.79 | 0.77 |
| 94 | 0.11 | 0.28 | 0.89 | 0.89 | -1.84 | -0.13 | 0.28 | 0.38 | 0.84 | 0.90 | 0.83 | 0.78 |

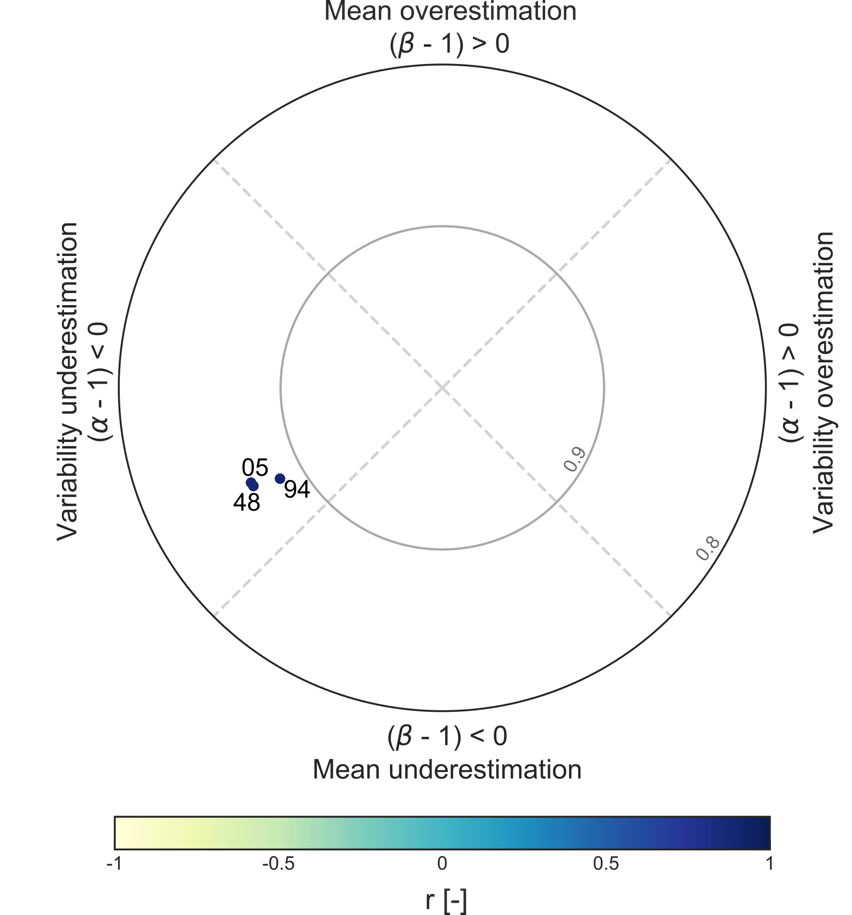


Figure A4: Polar plot of KGE for real case example