Technical note: Diagnostic efficiency – insights into model performance

Robin Schwemmle1, Dominic Demand1, Andreas Hartmann1, Markus Weiler1

1University of Freiburg, Institute of Earth and Environmental Sciences, 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

* 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

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. Component are entirely based on the flow duration curve. To include the missing temporal dimension we added a third component.

The first component reflects the input data 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 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 input data error and the model error remains.

The third metric component is the Pearson correlation r for the simulated and observed discharge time series.

Calculating the distance between , and r results in the Diagnostic efficiency (DE, Eq. (5)):

, (5)

However, the calculation of DE does not allow a diagnosis. Thus, we project DE in the 2-D space.

, (6)

, (7)

, (8)

,

,

,

,

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

, (9)

where …

, (10)

where …

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

, (11)

where ..

# 3 Proof of concept

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

Two sources of errors…

Mimicking model 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 input data errors:

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

Temporal mismatch due to model errors and/or input data errors:

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

Combination of model errors and input data errors:

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 model errors, input data errors and temporal mismatch:

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

# 4 Conclusions

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

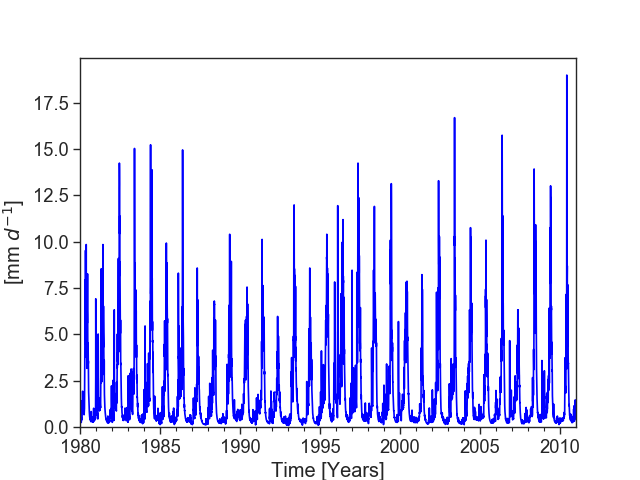


Figure 1: Observed streamflow time series

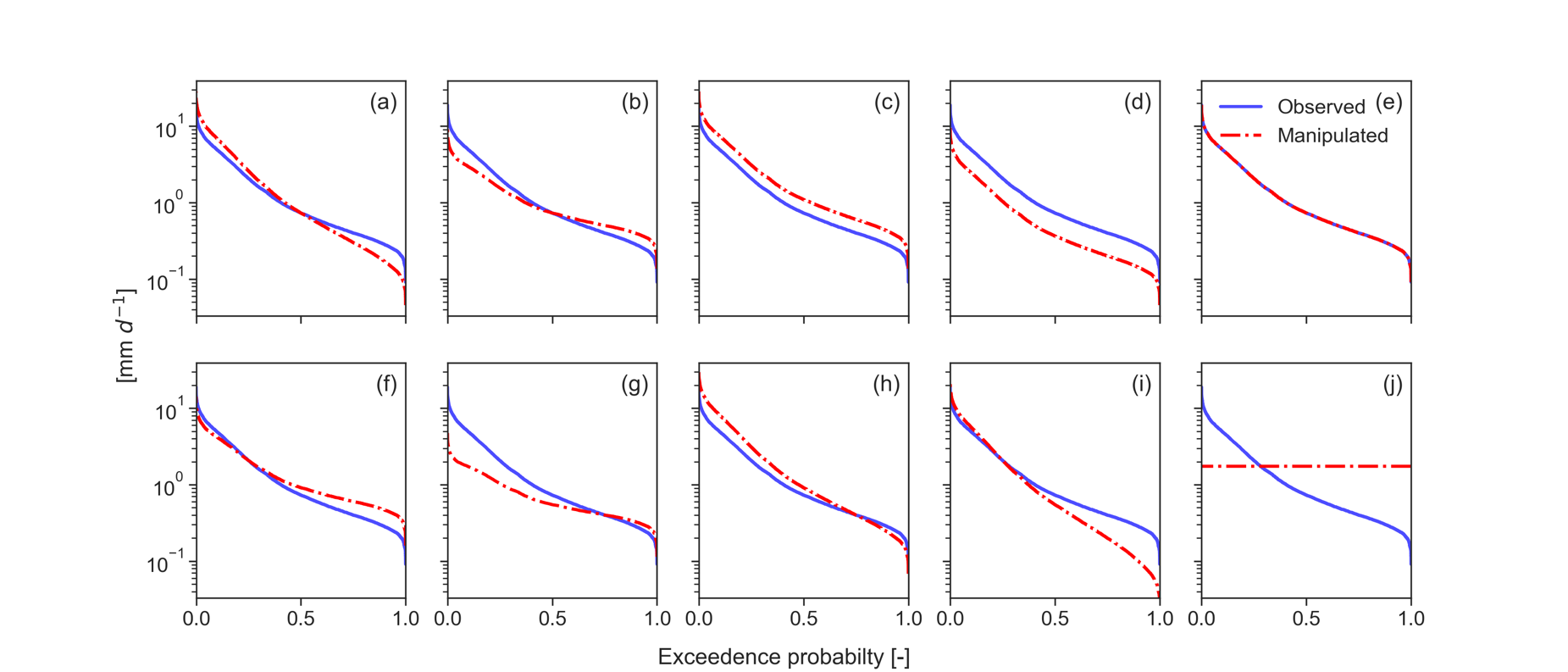


Figure 2: Flow duration curves of observed and manipulated streamflow time series

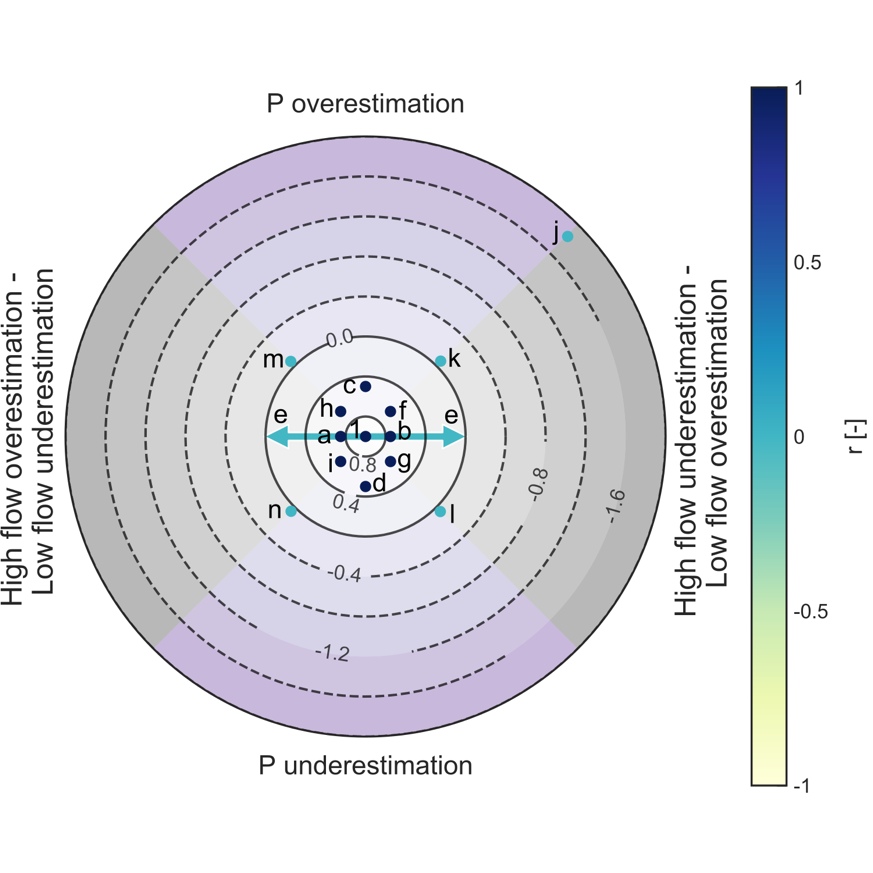


Figure 3: Diagnostic plot and mimicked errors

Table 1: DE, KGE and NSE for mimicked errors

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 1' | a | b | c | d | e | f | g | h | i | j | k | l | m | n |
| DE | 1 | 0.75 | 0.75 | 0.5 | 0.5 | 0.01 | 0.65 | 0.65 | 0.65 | 0.65 | -1.84 | -0.05 | -0.06 | -0.07 | -0.06 |
| KGE | 1 | 0.43 | 0.43 | 0.29 | 0.29 | 0.01 | 0.75 | 0.08 | 0.08 | 0.75 | -0.41 | -0.02 | -0.36 | -0.37 | -0.03 |
| NSE | 1 | 0.7 | 0.7 | 0.6 | 0.6 | -0.99 | 0.94 | 0.27 | 0.27 | 0.94 | 0 | -0.55 | -0.24 | -3.27 | -1.55 |

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

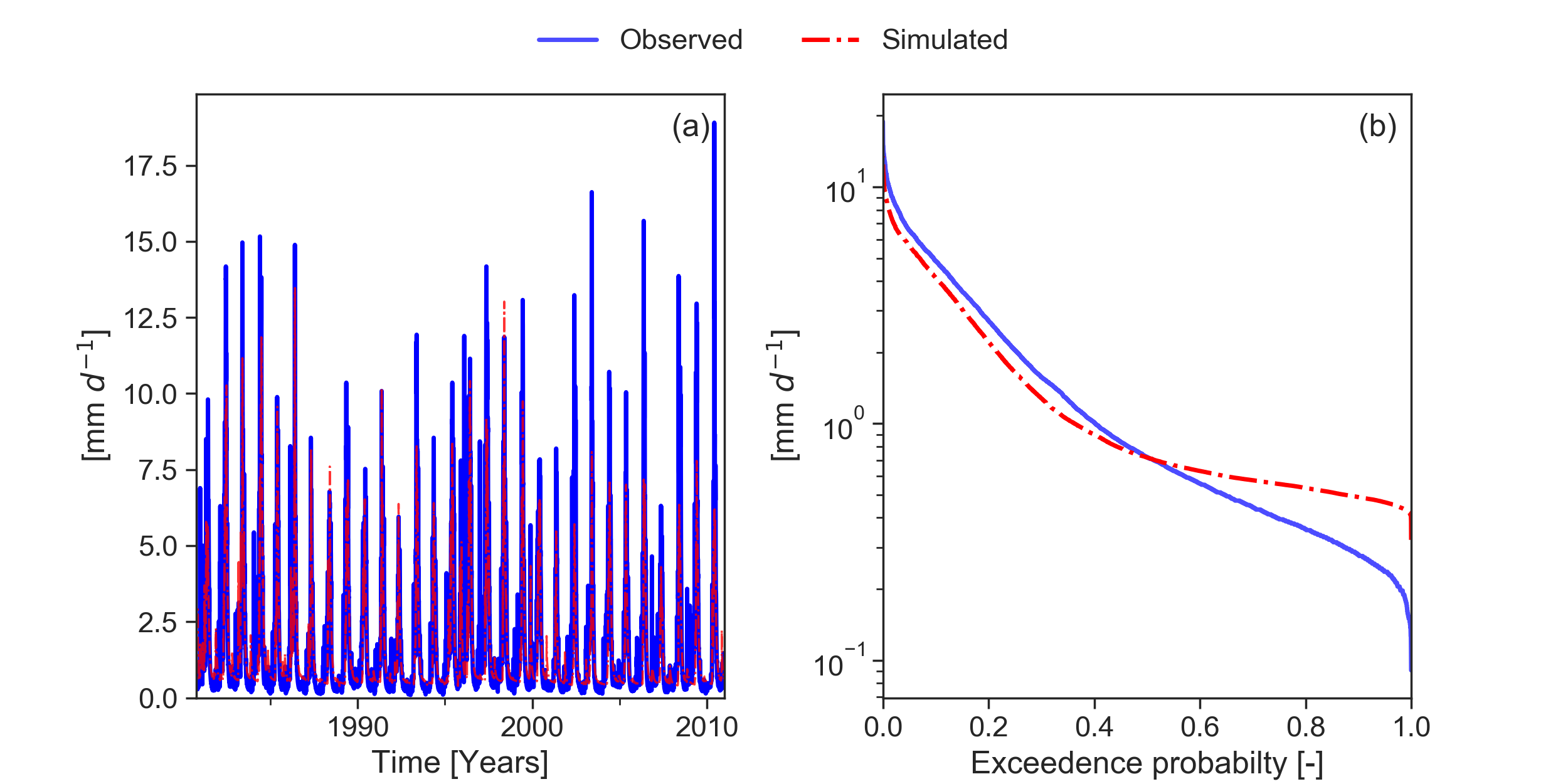


Figure 4: Simulated and observed streamflow of real case example (a) and the related flow duration curves (b)

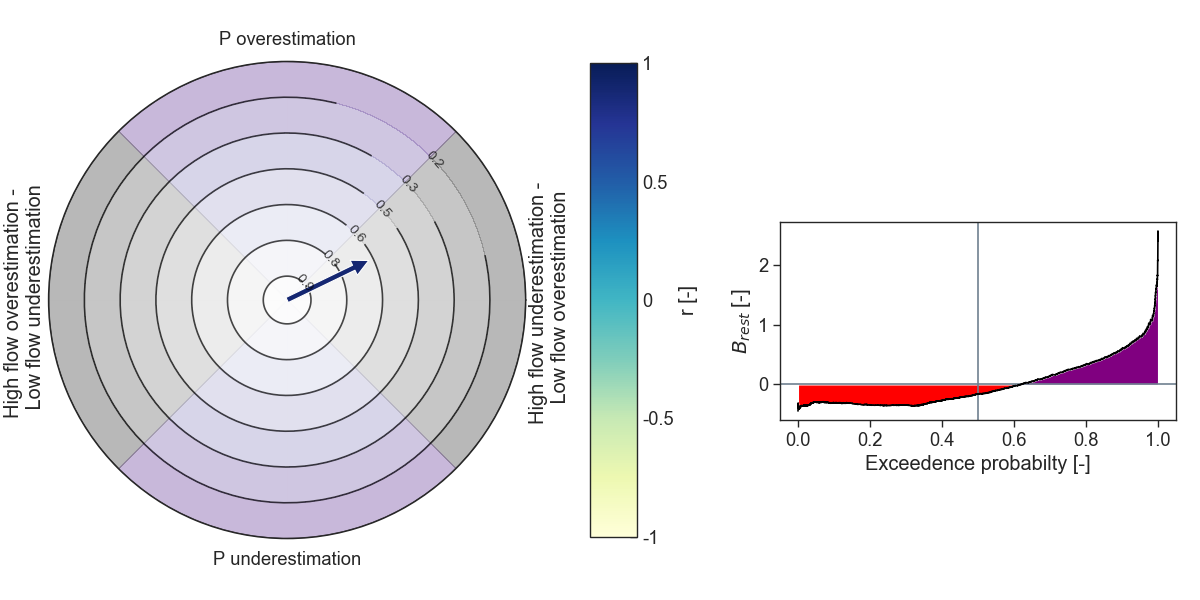


Figure 5: Diagnostic plot for real case example

# References

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.

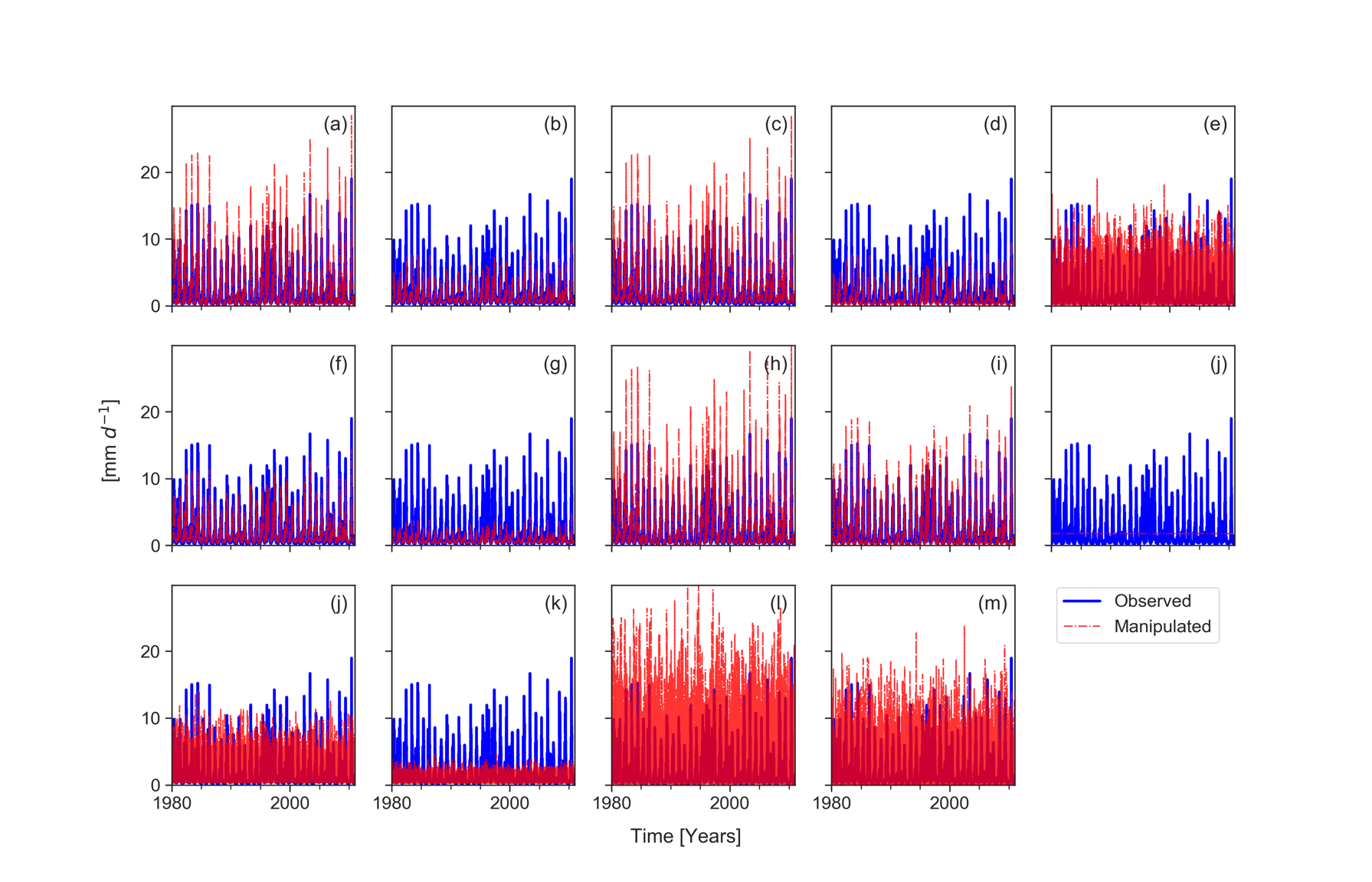


Figure A1: Time series of observed and manipulated streamflow

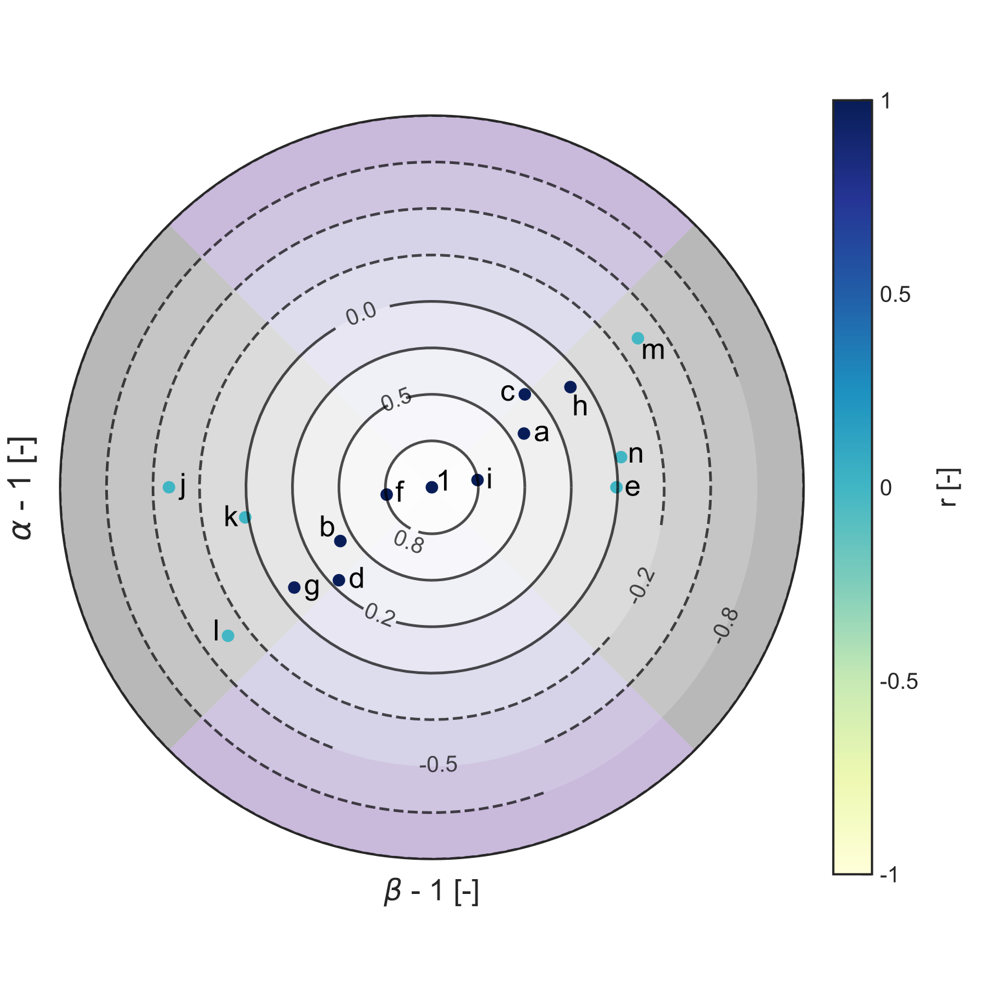


Figure A2: Polar plot of KGE

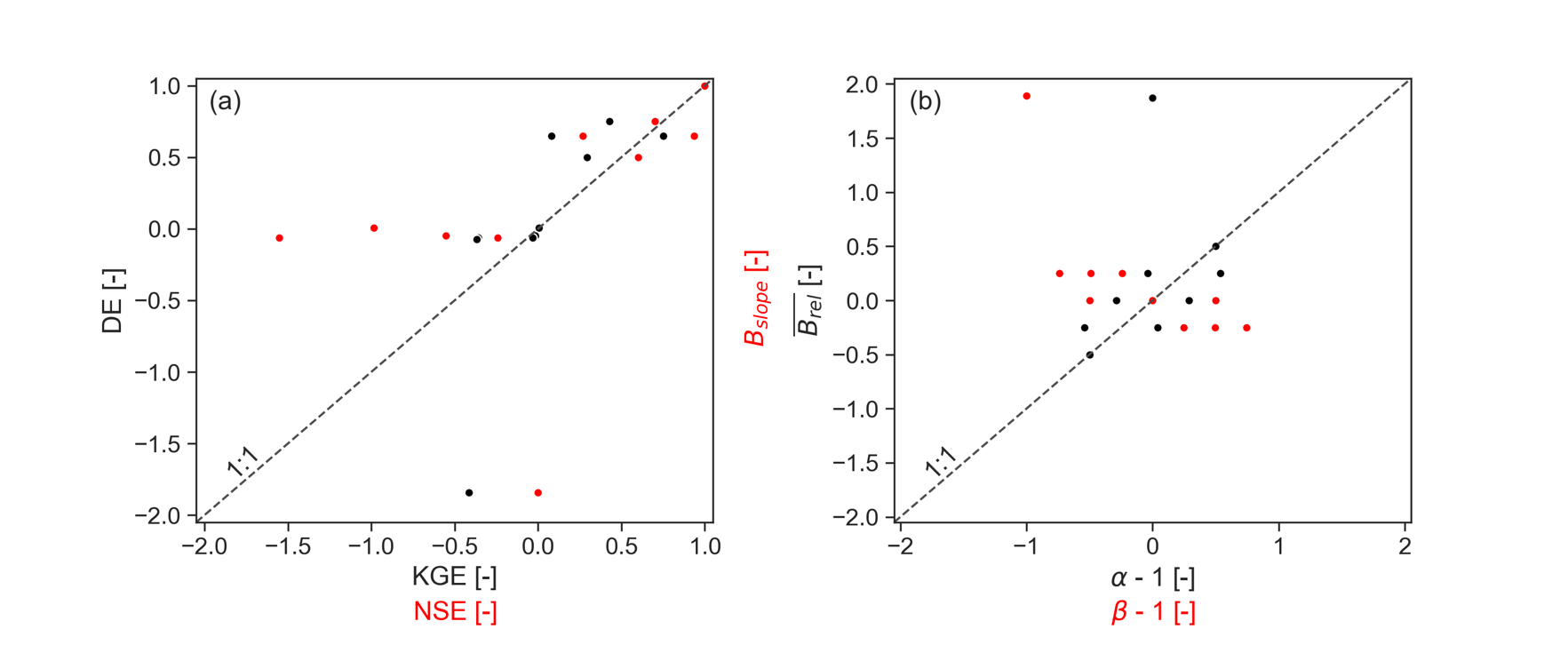


Figure A3: Scatterplot

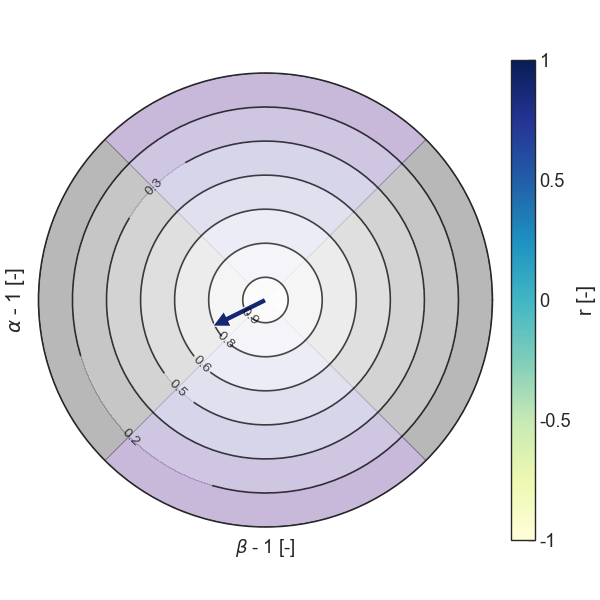


Figure A4: Polar plot of KGE for real case example