Technical note: Diagnostic efficiency – specific evaluation of 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
* Model uncertainty

Elaborate on well-established efficiency measures (Schaefli and Gupta, 2007)

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 – why?

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

Maybe we also need to discuss the ideas of signatures? (Gupta et al., 2008)

Or other approaches that tried to use diagnostic measures (e.g. Pechlivanidis, I. G., Jackson, B., & McMillan, H. (2010). The use of entropy as a model diagnostic in rainfall-runoff modelling.)

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 (but we are doing more, not only FDC)

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

Need for diagnostic approaches

# 2 Methodology

Errors in hydrological simulations may be caused by the following origins:

* 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)
* initial and boundary conditions

In general, the quality of observations should be verified before simulations are compared to. Observations which accuracy is not sufficient enough should not be considered for model evaluation. In order to reveal the origin of the errors we define three error types which the upper three error origins in the above mentioned list may be linked to:

* constant error (e.g. caused by consistently overestimated precipitation data)
* dynamic error (e.g. caused by storage routine)
* timing error (e.g. caused by model parameters)

In order to contribute to existing diagnostic evaluation approaches we introduce the non-normalized Diagnostic efficiency (*DEnn*, Eq. (1)):

, (1)

where is a measure for constant error, *|Barea|* for dynamic error, and *r* for timing error. Similar to *NSE* and *KGE*, *DE* ranges from 1 to -∞. *DE* = 1 indicates perfect agreement between simulations and observations.

First we introduce the three components which build up *DE*. The first two components and are based on the flow duration curve (FDC). To include the missing temporal dimension, we added a third component (*r*).

reflects the constant error and is represented by arithmetic mean of the relative bias (Eq. (2)):

, (2)

*i* represents the exceedance probability, *N* the total number of data points and *Brel* is the relative bias of the simulated and observed flow duration curve; = 0 indicates no constant error; < 0 indicates negative constant error; > 0 indicates positive constant error. The relative bias between the simulated and observed flow duration curve (*Brel*) calculates as follows (Eq. (4)):

, (4)

*Qsim* is the simulated streamflow at exceedance probability *i* and *Qobs* the observed streamflow at exceedance probability *i*.

The dynamic error which is described by the absolute area of the residual bias (*|Barea|*; Eq. (5)):

, (5)

where the residual bias *Brest* is integrated over the entire domain of the flow duration curve. Eq. (6) is inserted in Eq. (5):

, (6)

by subtracting we remove the constant error and the dynamic error remains. *|Barea|* = 0 indicates no dynamic error; *|Barea| > 0* indicates a dynamic error.

To consider timing errors the linear correlation between simulations and observations (r) is calculated (Eq. (7)):

(7)

where *Qsim* is the simulated streamflow at time *t*, *Qobs* the observed streamflow at time *t*, *μobs* the simulated mean streamflow, and *μobs* the observed mean streamflow.

* Discuss benchmark (Seibert et al., 2018)

Since the mean flow benchmark of *DE* (*DEmfb*; *Qsim (t) = μobs*) strongly depends on *Qobs (t)* (i.e. *DEmfb* is not constant across different time series), we scale *DE* by its *DEmfb* (*DE*; Eq. (8)):

, (8)

as such *DE* is scaled… The normalization is required to enable an inter-comparison of *DE* (e.g. for comparative hydrology).

*DE* can be used as other efficiency measures in simply optimizing for the highest value. However, the calculation of *DE* as it is does not allow a diagnosis. Thus, we project *DE* in a radial plane (i.e. similar to a clock). For this, we calculate the direction of the dynamic error (*Bdir*; Eq. (9)):

, (9)

where the integral of *Brest* includes values from 0th percentile to 50th percentile. (Explain why we only use the left half?)

In order to differentiate the dynamic error type, the slope of the remaining bias (*Bslope*; Eq. (10)) is computed:

, (10)

*Bslope* = 0 expresses no dynamic error; *Bslope* < 0 indicates that there is a tendency of simulations to overestimate high flows and underestimate low flows while *Bslope* > 0 indicates a tendency of simulations to underestimate high flows and overestimate low flows.

We used the inverse tangent to derive the ratio between constant error and dynamic error in radians (*ϕ*; Eq. (11)):

, (11)

Here we introduce conditions for which a diagnosis can be drawn. We set a threshold value (*lim*) for which metric components deviate and insert it in Eq. (1):

, (12)

This results a threshold value for DEnn (*DElim-nn*) which is normalized by (Eq. (13))

, (13)

Finally, the following conditions describe whether a diagnosis can be drawn (Eq. (14)):

, (14)

with

,

In case conditions for a diagnosis are not fulfilled, no diagnosis can be drawn when of the following conditions are true (Eq. (15)):

, (15)

*DS* denotes a deficient simulation and *GS* a good simulation.

In order to allow a comparison to the commonly used Kling-Gupta Efficiency (*KGE*; Gupta et al., 2009) and Nash-Sutcliffe Efficiency (*NSE*; Nash and Sutcliffe, 1970), we present the corresponding equations. We used the original *KGE* proposed in Gupta et al. (2009):

, (16)

where *α* represents the flow variability error, *β* is the bias term and *r* shows the linear correlation between simulations and observations (Eq. (16)):

, (17)

where *σobs* is the standard deviation in observations, *σsim* the standard deviation in simulations, *μobs* the arithmetic mean of observations, and *μsim* the arithmetic mean of simulations.

We calculated the mean flow benchmark of Kling-Gupta Efficiency (*KGEmfb*) recently proposed by Knoben et al. (2019):

, (18)

and used the benchmark to calculate the skill score (*KGEskill score;* Eq. (19)):

, (19)

*KGEskill score* > 0 indicate a better performance than the mean flow whereas *KGEskill score* < 0 indicates that simulations perform worse than the mean flow benchmark.

Nash-Sutcliffe Efficiency (NSE; Nash and Sutcliffe, 1970) calculates as follows (Eq. (20)):

, (20)

where *T* is the total number of time steps, *Qsim (t)* the simulated streamflow at time *t*, *Qobs (t)* the observed streamflow at time *t* and *μobs.* NSE = 1 displays perfect fit between simulations and observations; NSE = 0 indicates that simulations performs equally well as the mean of the observations; NSE < 0 indicates that simulations perform worse than the mean of the observations.

# 3 Proof of concept

We provide a proof of concept for which we used an observed streamflow time series from the CAMELS data set (Newman et al., 2015). Note that for this any streamflow time series which comes from a near-natural catchment and has sufficiently long temporal record. In order to mimic model errors, we systematically manipulated the observed time series (Table).

## 3.1 Mimicking errors

Mimicking dynamic errors:

1. Increase high flows – Decrease low flows: Multiplying the observed FDC with a linearly interpolated vector (1.5, 1.49, …, 0.49, 0.5) which amplifies the extremes (see Fig. 2a). Note that the original temporal order is maintained.
2. Decrease high flows – Increase low flows: Multiplying the observed FDC with a linearly interpolated vector (0.5, 0.49, …, 1.49, 1.5) which moderates the extremes (see Fig. 2b). Note that the original temporal order is maintained.

Mimicking constant error:

1. Positive offset: Multiplying the observed time series with a constant > 1 (see Fig. 2c).
2. Negative offset: Multiplying the observed time series with a constant < 1 (see Fig. 2d).

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

1. Shuffling: Randomizing the order of the observed time series (see Fig. 2e).

Combination of dynamic error and constant error:

1. Decrease high flows – Increase low flows and negative offset (see Fig. 2f).
2. Decrease high flows – Increase low flows and positive offset (see Fig. 2g).
3. Increase high flows – Decrease low flows and negative offset (see Fig. 2h).
4. Increase high flows – Decrease low flows and positive offset (see Fig. 2i).

Benchmark against KGE and NSE:

1. Mean flow benchmark (see Fig. 2j).

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

- intra-catchment comparison does not require normalization

# 4 Discussion and 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

*Code availability.* We provide a Python package de which can be used to mimick errors, calculate DE and the corresponding metric components and to produce the diagnostic polar plots. The stable version can be installed via the Python Package Index (PyPI), and the current development version is available at https://github.com/schwemro/de.

*Data availability.* The observed and simulated streamflow time series are part of the CAMELS dataset (Newman et al., 2015). The data can be downloaded from https://ncar.github.io/hydrology/datasets/CAMELS\_timeseries.

*Author contributions.* RS had the idea. RS, DD and MW jointly developed and designed the methodology. RS developed the code, produced the figures and tables, and wrote the first draft of the manuscript. The manuscript was revised by DD and MW and edited by RS.

*Competing interests.* The authors declare that they have no conflict of interest.

*Acknowledgements.*

*Financial support.* This research has been supported by …

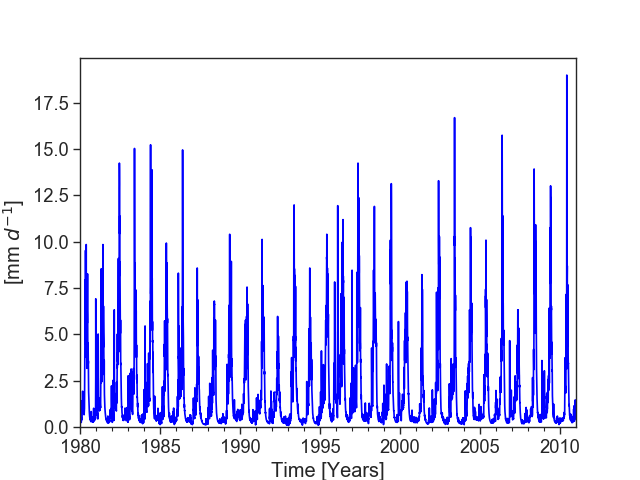


Figure 1: Observed streamflow time series from CAMELS dataset (Newman et al., 2015; gauge\_id: 13331500; gauge\_name: Minam River near Minam, OR, U.S.)

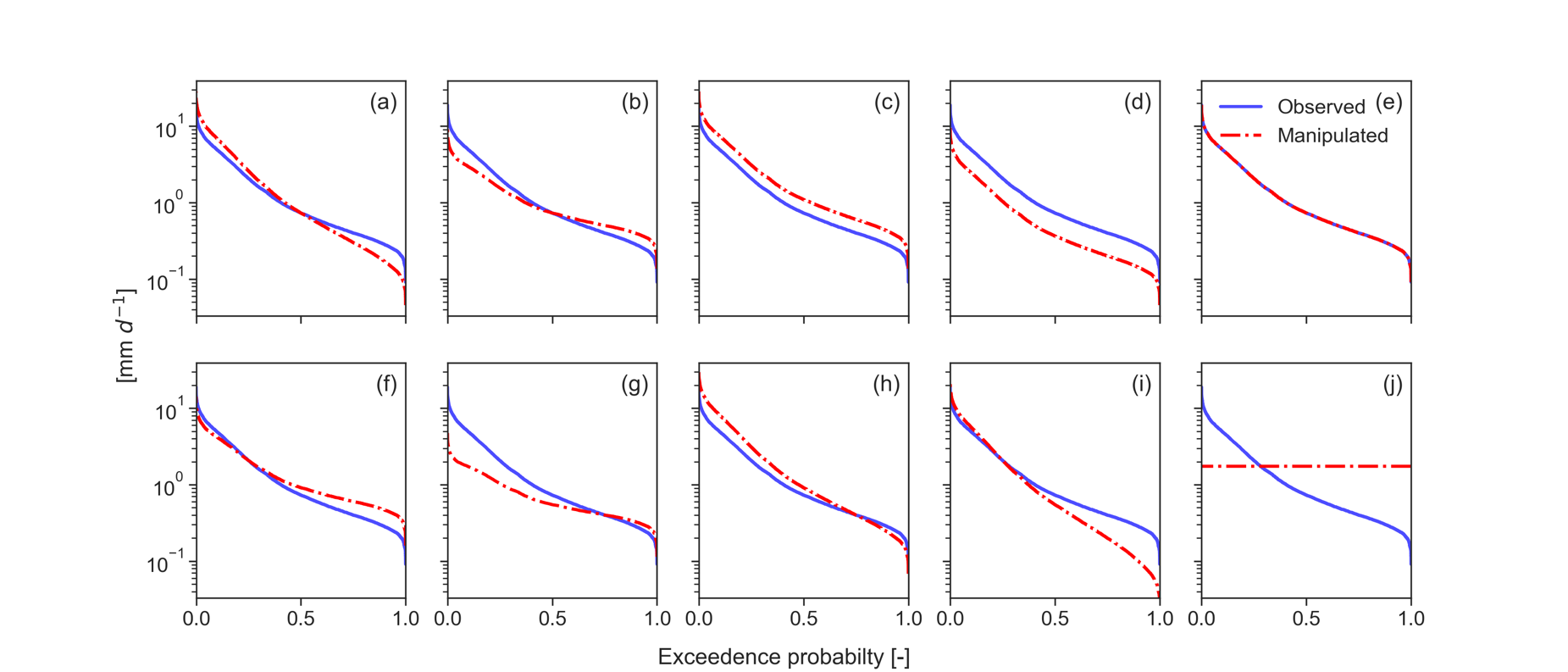


Figure 2: Flow duration curves (FDCs) of observed (blue) and manipulated (dashed red) streamflow time series. Manipulated FDCs are depicted for (a-b) dynamic errors only, (c-d) constant errors only, (e) timing error only; (f-i) combination of dynamic and constant error and (j) the mean flow benchmark.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

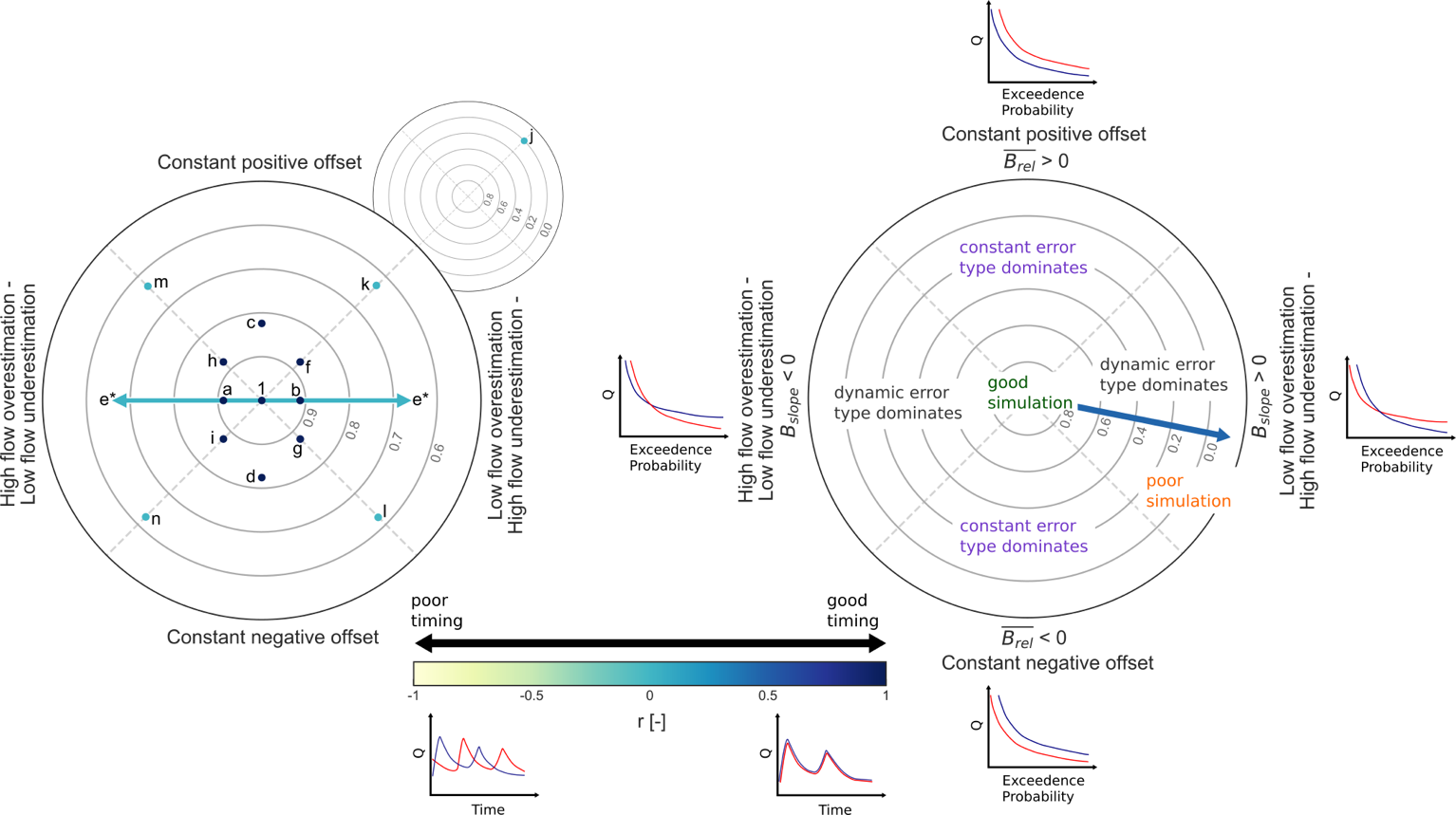


Figure 3: (left) Diagnostic polar plot for the mimicked errors (a-i; k-n) visualizing the overall model performance (*DE*; contour lines) and contribution of constant error, dynamic error and timing error(blue (yellow) indicates temporal match (mismatch)). (e\*) type of dynamic error cannot be distinguished. Inset shows diagnostic polar plot for the mean flow benchmark (j). (right) Annotated diagnostic polar plot illustrating the interpretation (similar to Zipper et al. (2018)). Hypothetic FDC plots and hydrograph plots give examples for the error types.

Table 1: Comparison of *DE*, *KGEskill score* 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 |
| *KGEskill score* | 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
* Best KGE values and NSE values show similar tendency, but not DE (Table 1: f and i)

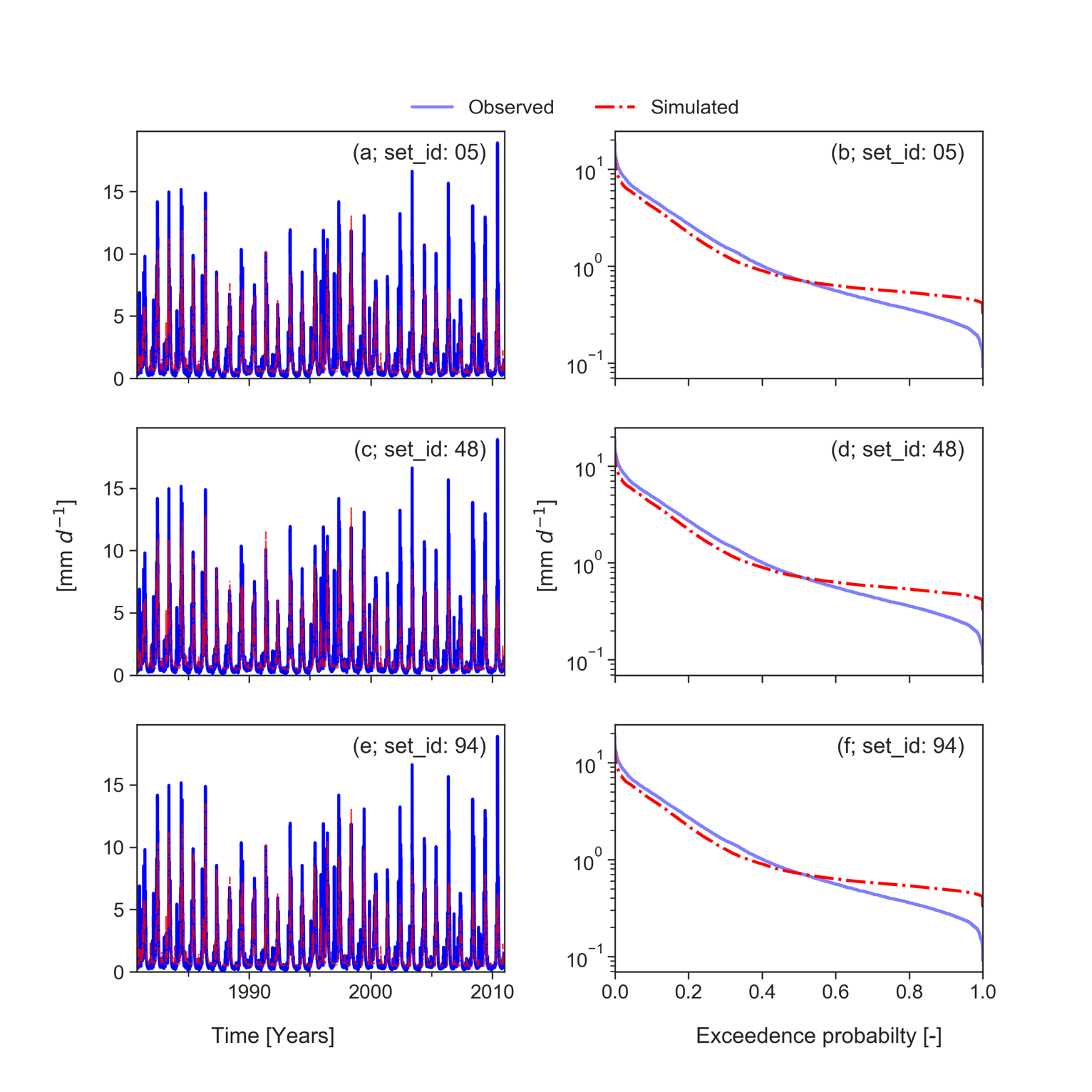


Figure 4: Simulated and observed streamflow time series of real case example (a, c and e) and the related flow duration curves (b, d and f). Time series are derived from the CAMELS dataset (Newman et al., 2015). Observed time series is the same as in Figure 1. Simulated time series had been produced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

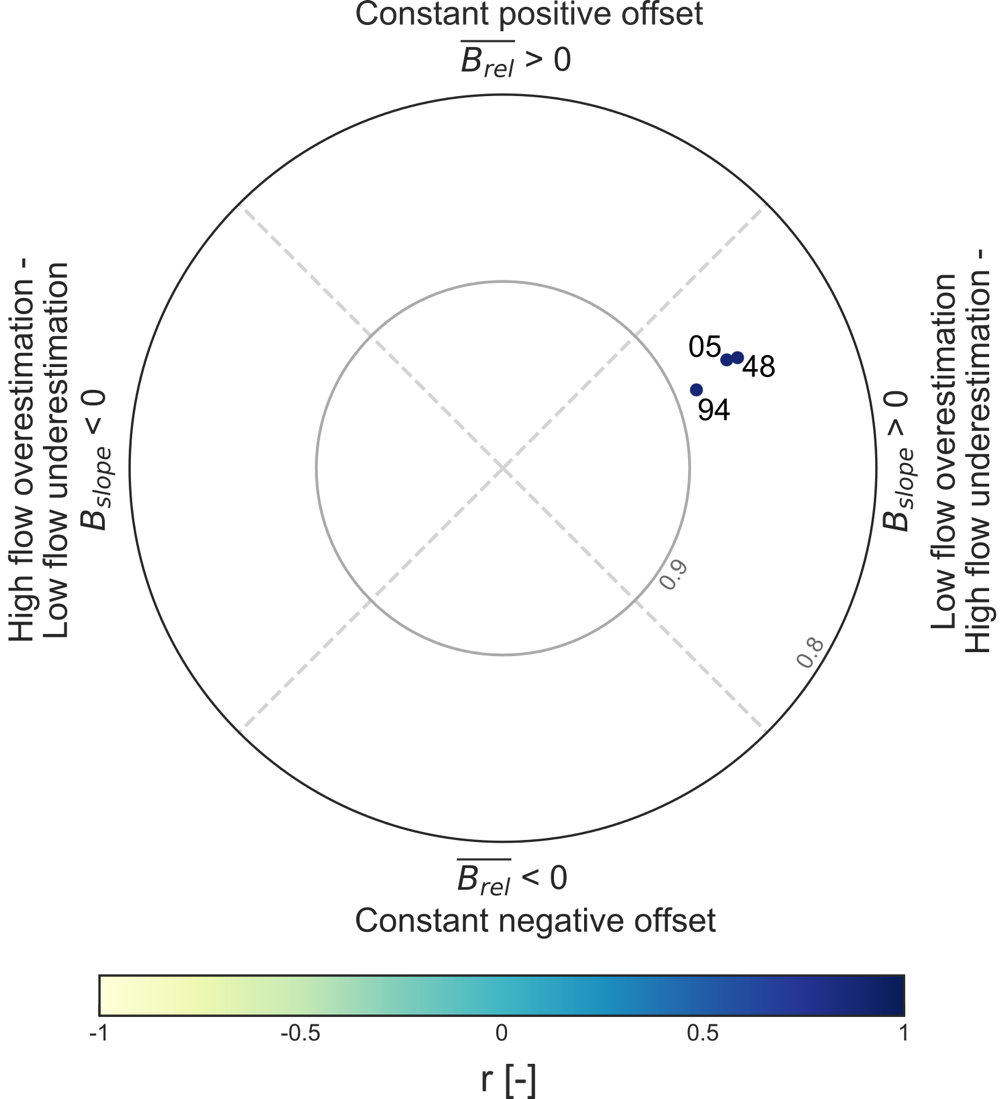


Figure 5: Diagnostic plot for real case example. Three different simulation runs are evaluated (05, 48, 94; see Figure 4). All simulations perform well. However, the remaining error is dominated by the dynamic error type while timing is excellent.

* 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., Fowler, K. J. A., Peel, M. C., and Woods, R. A.: Modular Assessment of Rainfall–Runoff Models Toolbox (MARRMoT) v1.2: an open-source, extendable framework providing implementations of 46 conceptual hydrologic models as continuous state-space formulations, Geosci. Model Dev., 12, 2463-2480, 10.5194/gmd-12-2463-2019, 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.

Seibert, J., Vis, M. J. P., Lewis, E., and van Meerveld, H. J.: Upper and lower benchmarks in hydrological modelling, Hydrological Processes, 32, 1120-1125, 10.1002/hyp.11476, 2018.

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.

Zipper, S. C., Dallemagne, T., Gleeson, T., Boerman, T. C., and Hartmann, A.: Groundwater Pumping Impacts on Real Stream Networks: Testing the Performance of Simple Management Tools, Water Resources Research, 54, 5471-5486, 10.1029/2018wr022707, 2018.

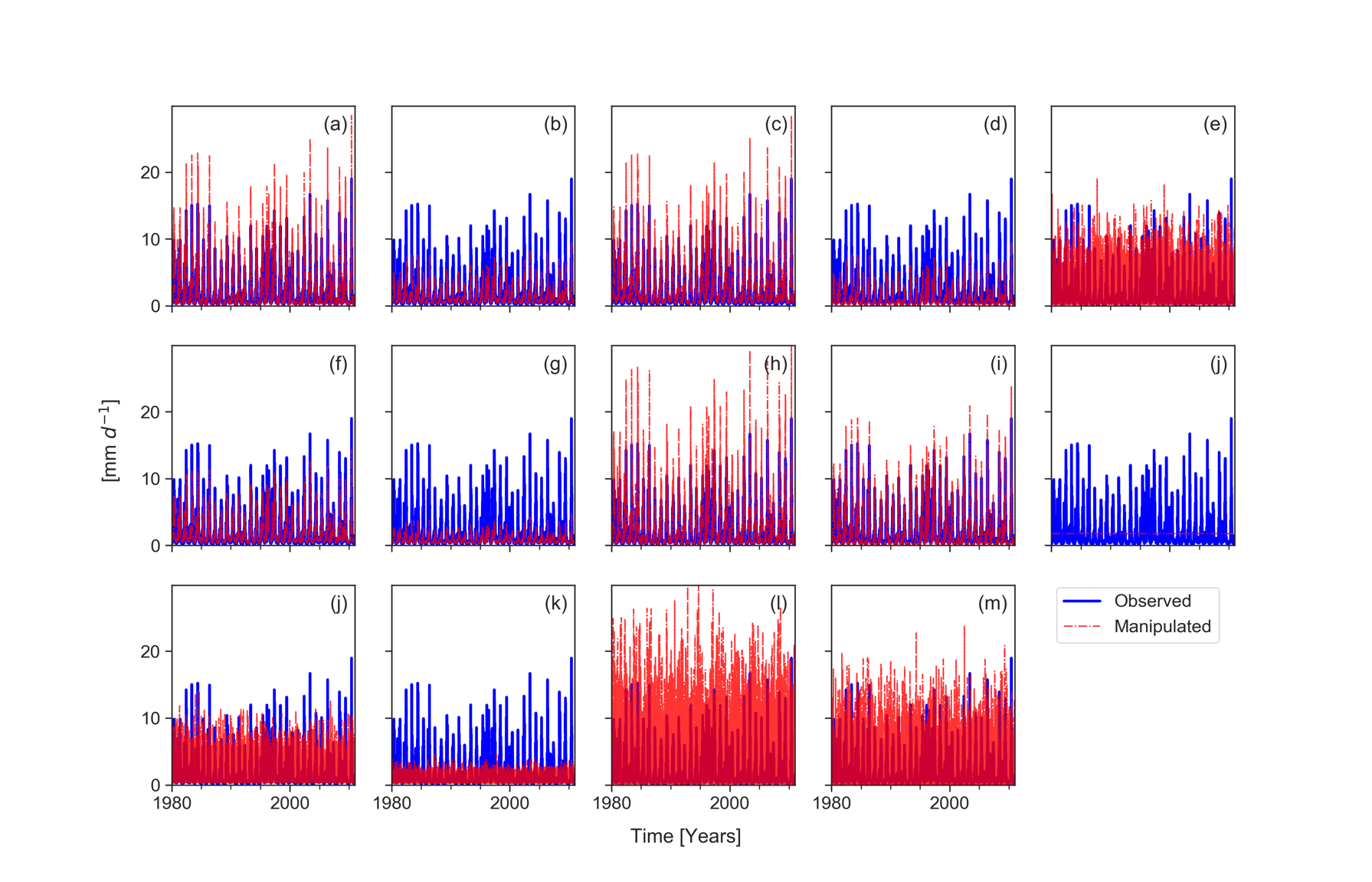


Figure A1: Time series of observed and manipulated streamflow

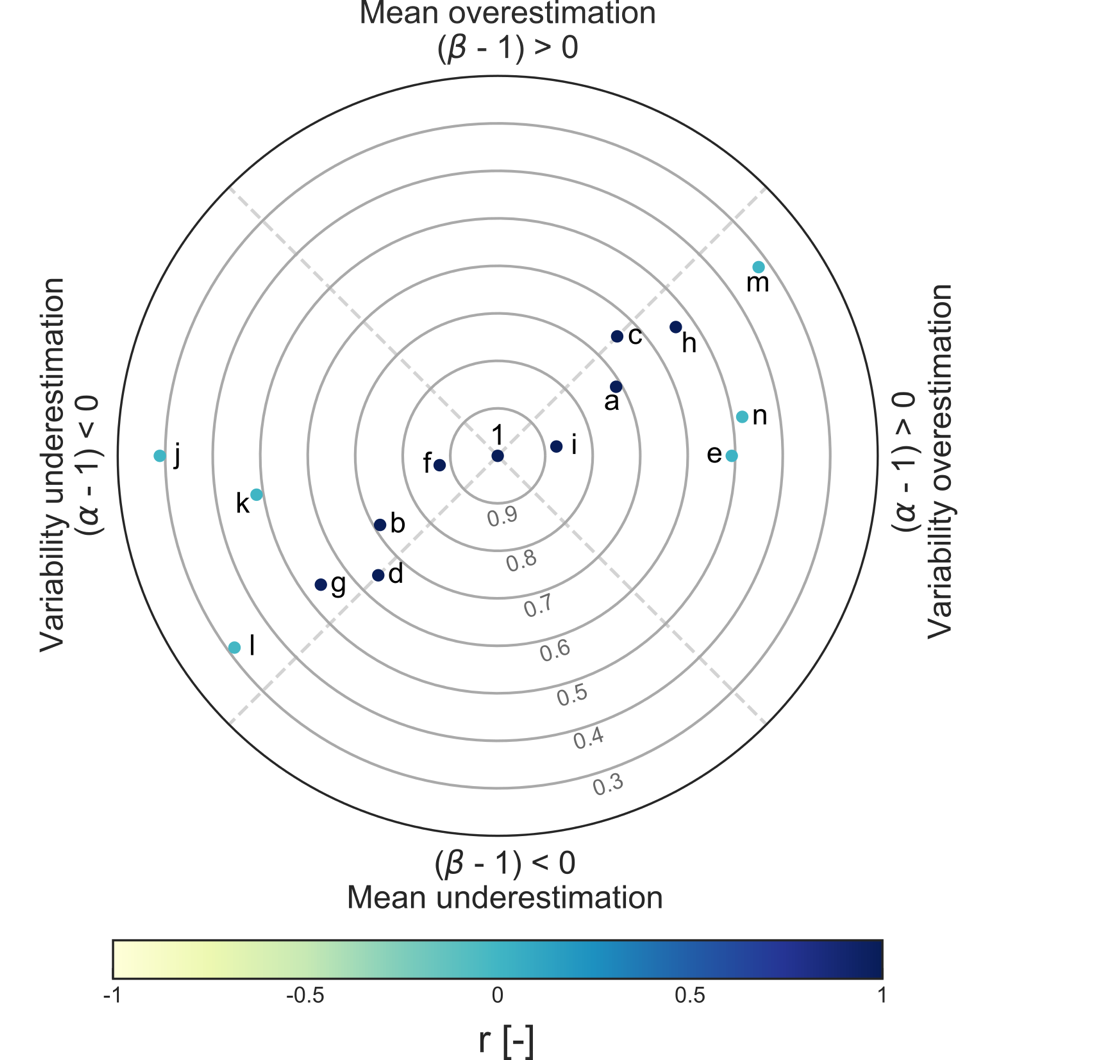


Figure A2: Polar plot of *KGEskill score*

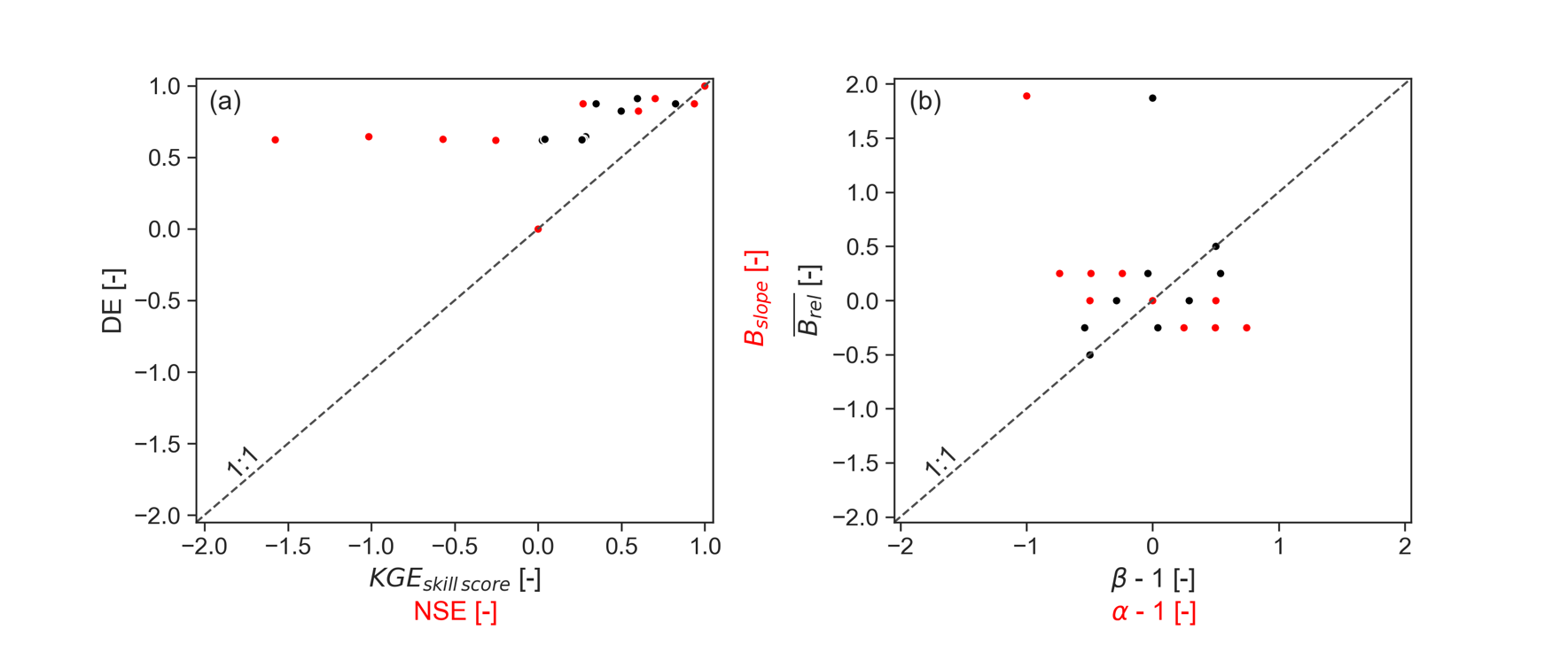


Figure A3: (a) Scatterplot to compare DE with KGE (black) and DE with NSE (red), respectively. (b) Scatterplot to compare with (black) and with (red), respectively.

Table A1: Comparison of metric components for mimicked errors

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 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 |
| *|Barea|* | 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 |
| *Bdir* | 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 |
| *Bslope* | -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 |
| α | 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 |
| β | 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 |  | *|Barea|* | *r* | *DE* | *DEmfb* | *Bdir* | *Bslope* | ϕ | *KGEskill score* | β | α | *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 |

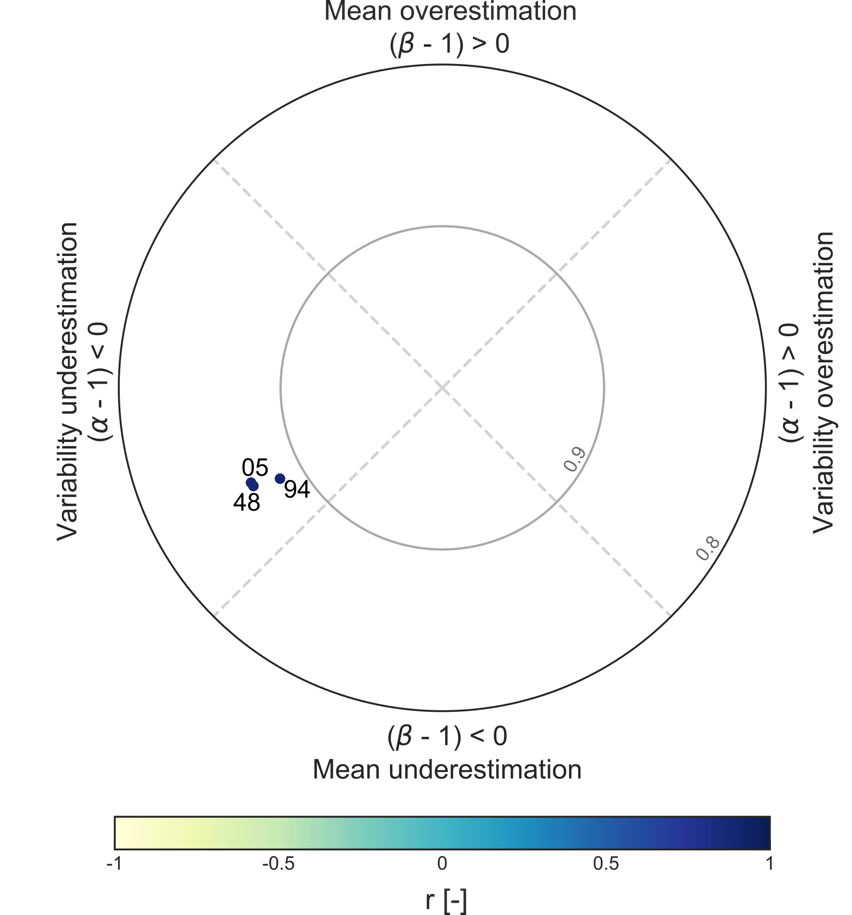


Figure A4: Polar plot of *KGEskill score*for real case example