

Supplementary Material for "Metrics that Matter: Objective Functions and Their Impact on Signature Representation in Conceptual Hydrological Models"

Peter Wagener et al.

S1.1 List of MARRMoT Models

Table S1: MARRMoT v2.1 models (1/2)

Number	Name	Storages	Params
01	collie1	1	1
02	wetland	1	4
03	collie2	1	4
04	newzealand1	1	6
05	ihacres	1	7
06	alpine1	2	4
07	gr4j	2	4
08	us1	2	5
09	susannah1	2	6
10	susannah2	2	6
11	collie3	2	6
12	alpine2	2	6
13	hillslope	2	7
14	topmodel	2	7
15	plateau	2	8
16	newzealand2	2	8
17	penman	3	4
18	simhyd	3	7

Continued on next page

Table S1: MARRMoT v2.1 models (2/2)

Number	Name	Storages	Params
25	tcm	4	6
26	flexi	4	10
27	tank	4	12
28	xinjiang	4	12
29	hymod	5	5
30	mopex2	5	7
31	mopex3	5	8
32	mopex4	5	10
33	sacramento	5	11
34	flexis	5	12
35	mopex5	5	12
36	modhydrolog	5	15
37	hbv	5	15
38	tank2	5	16
39	mcrm	5	16
40	smar	6	8
41	nam	6	10
42	hycymodel	6	12

Continued on next page

Table S1: MARRMoT v2.1 models (1/2) (Continued)

Number	Name	Storages	Params
19	australia	3	8
20	gsfb	3	8
21	flexb	3	9
22	vic	3	10
23	lascam	3	24
24	mopex1	4	5

Table S1: MARRMoT v2.1 models (2/2) (Continued)

Number	Name	Storages	Params
43	gsmsocnt	6	12
44	echo	6	16
45	prms	7	18
46	classic	8	12
47	IHM19	4	16

S1.2 Catchment Clustering

To select these basins, we categorized the 2901 available CARAVAN catchments by using a climate indices approach as applied in Knoben et al. (2018). It is based on a monthly moisture index (Willmott and Feddema, 1992) which is defined as follows:

$$MI = \begin{cases} 1 - \frac{E_p(t)}{P(t)}, & P(t) > E_p(t) \\ 0, & P(t) = E_p(t) \\ \frac{P(t)}{E_p(t)} - 1, & P(t) < E_p(t) \end{cases} \quad (1)$$

Where $E_p(t)$ is monthly values of potential evapotranspiration and $P(t)$ is monthly precipitation.

Based on this index, the dimensionless indices of aridity (I_M) and seasonality ($I_{M,r}$) can be derived:

$$I_M[-1, 1] = \frac{1}{12} \sum_{t=1}^{t=12} MI(t) \quad (2)$$

The moisture index I_M indicates the mean aridity of the location, which negative values indicating water limitation and positive values energy limitation.

$$I_{M,r}[0, 2] = \max(MI) - \min(MI) \quad (3)$$

The seasonality index $I_{M,r}$ shows the variance in moisture availability, with higher values indicating more variance.

Finally, the fraction of snow f_s is calculated using a threshold temperature of $T_0 = 0^\circ C$ in our case:

$$f_s = \frac{\sum P(T(t) \leq T_0)}{\sum_{t=1}^{t=12} P(t)} \quad (4)$$

f_s adds additional information regarding the fraction of precipitation that falls as snow and ranges from 0 (none) to 1 (all).

We used a k-means clustering algorithm to assign each of the 2901 catchments to one of ten clusters according to their aridity, seasonality and fraction of snow. We then picked one catchment closest to the centre of each cluster. Figure S2 shows the ten obtained clusters as well as the cluster location of the catchments that were selected (purple dots). When selecting the catchments we considered two things. First, we limited the catchment size from 100 to 1000 km^2 to increase the probability of an adequate signature representation through the calibrated models. Smaller catchments are less likely to dampen catchment responses and therefore better contain scale-dependent effects that also affect signature representation. Second, we ensured that the catchments had an data length of at least 20 years. To ensure the reliability of the identified clusters we repeated the clustering ten times with random seeds and could not identify large changes in the location of the centroids.

K-means clustered CARAVAN catchments in climate index space

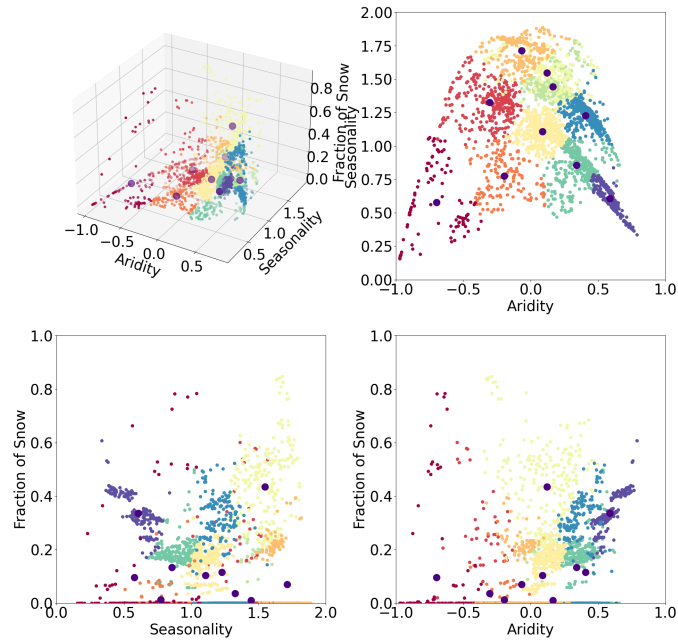


Figure S2. Results of the k-means clustering of 2901 CARAVAN catchments for their aridity, seasonality and fraction of snow. Each colored point cloud refers to one cluster and the thicker purple dots are the catchments selected for this study.

S1.3 Difference Calibration and Validation

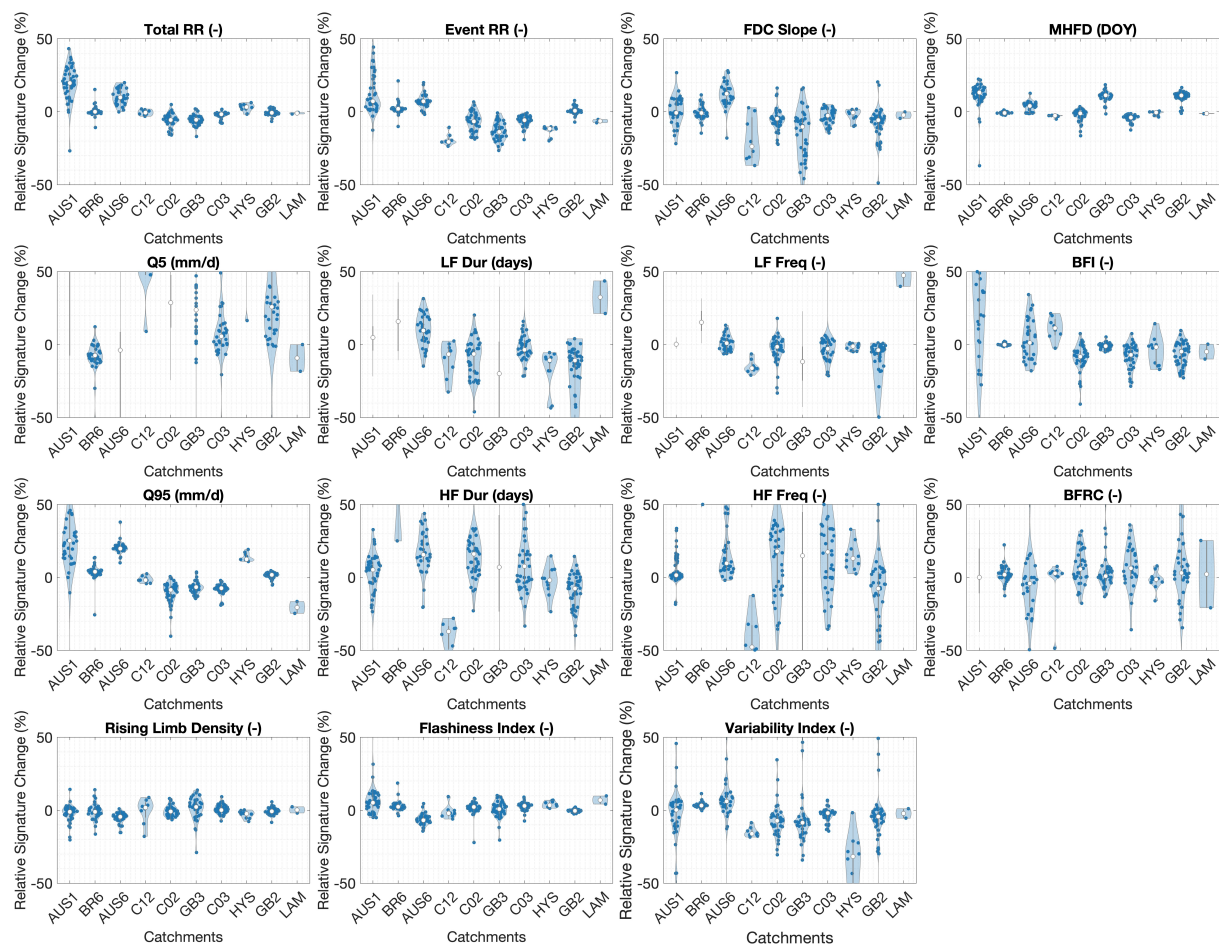


Figure S3. Calibration vs. Evaluation Differences

S1.4 Test Cases Error Metrics

The motivation behind the used error metric was to quantify the signature error in a manner that is able to capture both the deviation from the observed value and the performance relative to the alternative objective functions while doing so for a large numbers of models. Additionally, we would like to be the metric representative when compared across signatures.

To achieve this, we first calculate the difference between simulated and observed value for all models, catchments, signatures and objective functions. In a first step of aggregation, we calculate the median over the models (which exceeded the benchmark), denoted through subscript i . This is a robust way to find a representative values. Next, we normalize by absolute of the maximum value found for this signature across all objective functions and per catchment. This lead to the worst performance being characterized by ± 1 for each catchment. This can further be aggregated by calculating the median over all catchments (denoted through j) leading to a single error metric per signature and objective function.

This following section clarifies the applied metrics using a generic example. Each Matrix represents the signature values depending on the catchment (columns) and model (rows) for a specific objective function.

$$\text{Matrix 1 (Objective Function 1): } \begin{bmatrix} 0.80 & 0.40 & 0.60 \\ 0.75 & 0.38 & 0.62 \\ 0.70 & 0.35 & 0.58 \end{bmatrix}$$

$$\text{Matrix 2 (Objective Function 2): } \begin{bmatrix} 0.60 & 0.30 & 0.52 \\ 0.55 & 0.28 & 0.51 \\ 0.50 & 0.32 & 0.50 \end{bmatrix}$$

$$\text{Matrix 3 (Objective Function 3): } \begin{bmatrix} 0.70 & 0.36 & 0.44 \\ 0.66 & 0.34 & 0.45 \\ 0.61 & 0.35 & 0.43 \end{bmatrix}$$

Error Metric:

$$EM_k = med_j(EM_{jk}) = med_j \left(\frac{med_i(S_{ijk} - y_j)}{|max(med_i(S_{ijk} - y_j))|} \right) \quad (5)$$

Observed Values:

$$\begin{bmatrix} 0.75 & 0.37 & 0.61 \end{bmatrix}$$

For the applied error metric, we calculate the error in signature representation for all of them:

$$\text{Error Matrix 1: } \begin{bmatrix} 0.05 & 0.03 & -0.01 \\ 0.00 & 0.01 & 0.01 \\ -0.05 & -0.02 & -0.03 \end{bmatrix}$$

$$\text{Error Matrix 2: } \begin{bmatrix} -0.15 & -0.07 & -0.09 \\ -0.20 & -0.09 & -0.10 \\ -0.25 & -0.05 & -0.11 \end{bmatrix}$$

$$\text{Error Matrix 3: } \begin{bmatrix} -0.05 & -0.01 & -0.17 \\ -0.09 & -0.03 & -0.16 \\ -0.14 & -0.02 & -0.18 \end{bmatrix}$$

Next, we can calculate the median over the all the modelled values:

$$\text{Median Error Matrix 1: } \begin{bmatrix} 0.00 & 0.01 & -0.01 \end{bmatrix}$$

$$\text{Median Error Matrix 2: } \begin{bmatrix} -0.20 & -0.07 & -0.10 \end{bmatrix}$$

$$\text{Median Error Matrix 3: } \begin{bmatrix} -0.09 & -0.02 & -0.17 \end{bmatrix}$$

To normalize, mainly to make the results comparable for multiple signatures with different ranges, we divide the error by the absolute of the maximum median error for each catchment. That means the range will be limited from -1 to +1 with an value of ± 1 indicating that the largest error was found for this objective function in this catchment. A value of 0 indicates that the median signature over all models equals the observed value.

$$\text{Absolute Maximum Median Error } \begin{bmatrix} 0.20 & 0.07 & 0.17 \end{bmatrix}$$

Based on this, we can calculate EM_{jk} for each catchment:

$$EM_{j1}: \begin{bmatrix} 0.00 & 0.14 & -0.06 \end{bmatrix}$$

$$EM_{j2}: \begin{bmatrix} -1.00 & -1.00 & -0.59 \end{bmatrix}$$

$$EM_{j3}: \begin{bmatrix} -0.45 & -0.29 & -1.00 \end{bmatrix}$$

For this test case, OF2 is worst metric regarding signature representation for catchments 1 and 2 while OF3 has worst representation in catchment 3. OF1 is the best metric for all of the catchments.

Finally, we can derive EM_k , the median normalized error for each objective function for this signature.

$$EM_1 = \begin{bmatrix} 0.00 \end{bmatrix}$$

$$EM_2 = [-1.00]$$

$$EM_3 = [-0.45]$$

This median is only useful for larger amounts of catchments, but the applied sample size (10) should be sufficient. Alternatively, one can use the mean as the metric is strictly contained in range.

S1.5 Violinplots for Signature Error

We can plot the derived errors for selected signatures to evaluate the impact of the objective function on the signature representation.

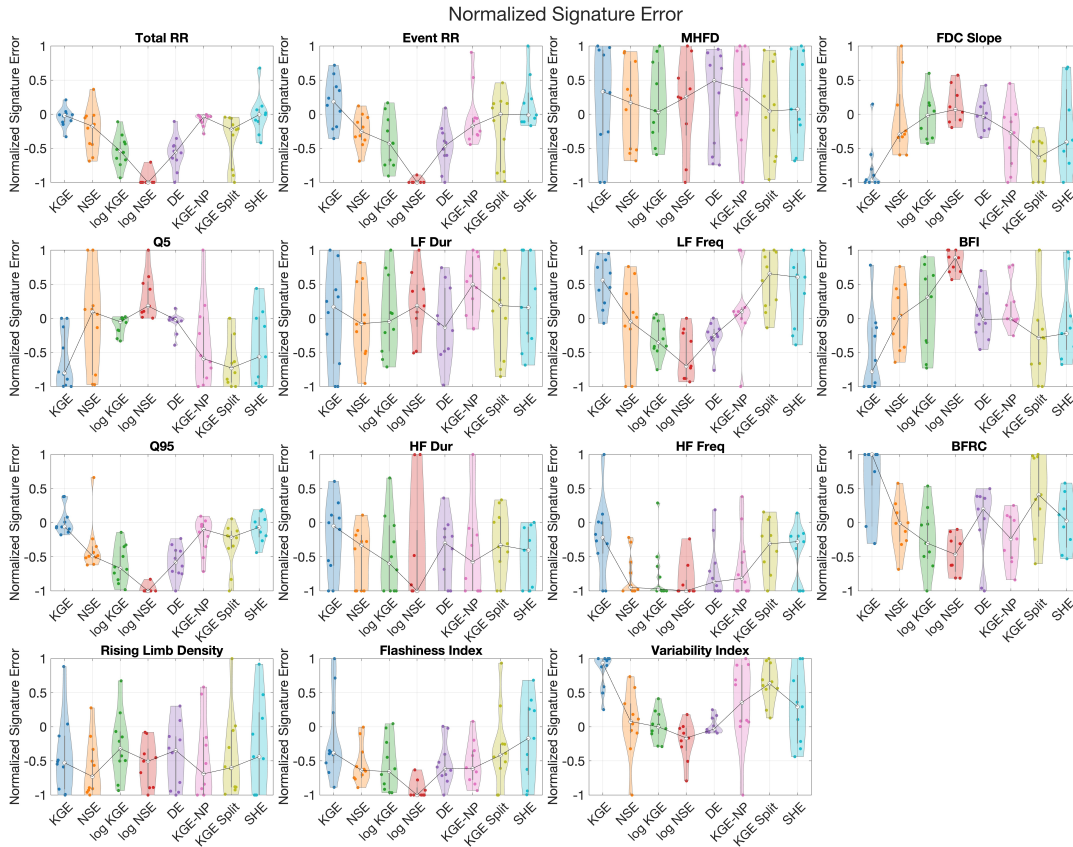


Figure S4. Variance-normalized Error over all Catchments and Models for 9 Signatures

The violinplots show the distribution of the median error over all models for each objective function and signature. This distribution represents the distribution for the objective function with each dot representing a catchment.

S1.6 Table with Normalized Error Values

Table S2. Medians and standard deviations of normalized signature errors for each objective function across 15 hydrological signatures.

Objective	Total RR (-)	Event RR (-)	MHFD (DOY)	FDC Slope (-)	Q5 (mm/d)	LF Dur (days)	LF Freq (-)	BFI (-)	Q95 (mm/d)	HF Dur (days)	HF Freq (-)	BFRC (-)	Rising Limb Density (-)	Flashiness Index (-)	Variability Index (-)
KGE median	-0.018	0.182	0.331	-0.961	-0.802	1.000	-0.050	-0.220	-0.384	0.168	0.559	-0.780	-0.060	-0.529	0.934
KGE sd	0.140	0.355	0.790	0.377	0.422	0.515	0.547	0.576	0.612	0.721	0.361	0.599	0.202	0.573	0.267
NSE median	-0.176	-0.251	0.170	-0.293	0.095	-0.010	-0.331	-0.949	-0.636	-0.078	-0.049	0.022	-0.489	-0.729	0.078
NSE sd	0.323	0.240	0.640	0.576	0.718	0.365	0.439	0.308	0.293	0.553	0.666	0.467	0.377	0.427	0.485
log-KGE median	-0.537	-0.425	0.034	-0.021	-0.050	-0.303	-0.596	-0.971	-0.658	-0.044	-0.353	0.309	-0.667	-0.319	-0.001
log-KGE sd	0.234	0.372	0.580	0.329	0.130	0.467	0.575	0.409	0.385	0.589	0.257	0.612	0.274	0.482	0.220
log-NSE median	-1.000	-1.000	0.253	0.069	0.181	-0.466	-1.000	-1.000	-1.000	0.183	-0.700	0.890	-1.000	-0.505	-0.164
log-NSE sd	0.099	0.037	0.696	0.281	0.337	0.293	0.859	0.265	0.132	0.496	0.359	0.163	0.057	0.338	0.292
Diagnostic Efficiency median	-0.554	-0.464	0.493	-0.028	-0.020	0.200	-0.299	-0.870	-0.615	-0.131	-0.250	-0.013	-0.580	-0.346	-0.033
Diagnostic Efficiency sd	0.251	0.323	0.731	0.235	0.149	0.575	0.477	0.417	0.322	0.552	0.214	0.363	0.235	0.470	0.120
KGE-np median	-0.039	-0.173	0.363	-0.271	-0.591	-0.237	-0.576	-0.817	-0.610	0.495	0.100	-0.008	-0.097	-0.690	0.356
KGE-np sd	0.090	0.432	0.661	0.478	0.634	0.372	0.793	0.490	0.312	0.439	0.563	0.346	0.261	0.608	0.613
KGE-split median	-0.221	0.000	0.047	-0.628	-0.727	0.408	-0.331	-0.309	-0.416	0.189	0.654	-0.287	-0.214	-0.599	0.633
KGE-split sd	0.376	0.520	0.680	0.306	0.397	0.588	0.478	0.436	0.581	0.677	0.428	0.717	0.334	0.617	0.291
SHE median	-0.006	-0.010	0.073	-0.407	-0.564	0.027	-0.404	-0.278	-0.169	0.160	0.604	-0.221	-0.067	-0.437	0.294
SHE sd	0.309	0.390	0.683	0.624	0.552	0.410	0.456	0.433	0.606	0.679	0.536	0.596	0.206	0.689	0.542

S1.7 Additional Plots for Detailed Analysis

Event RR:

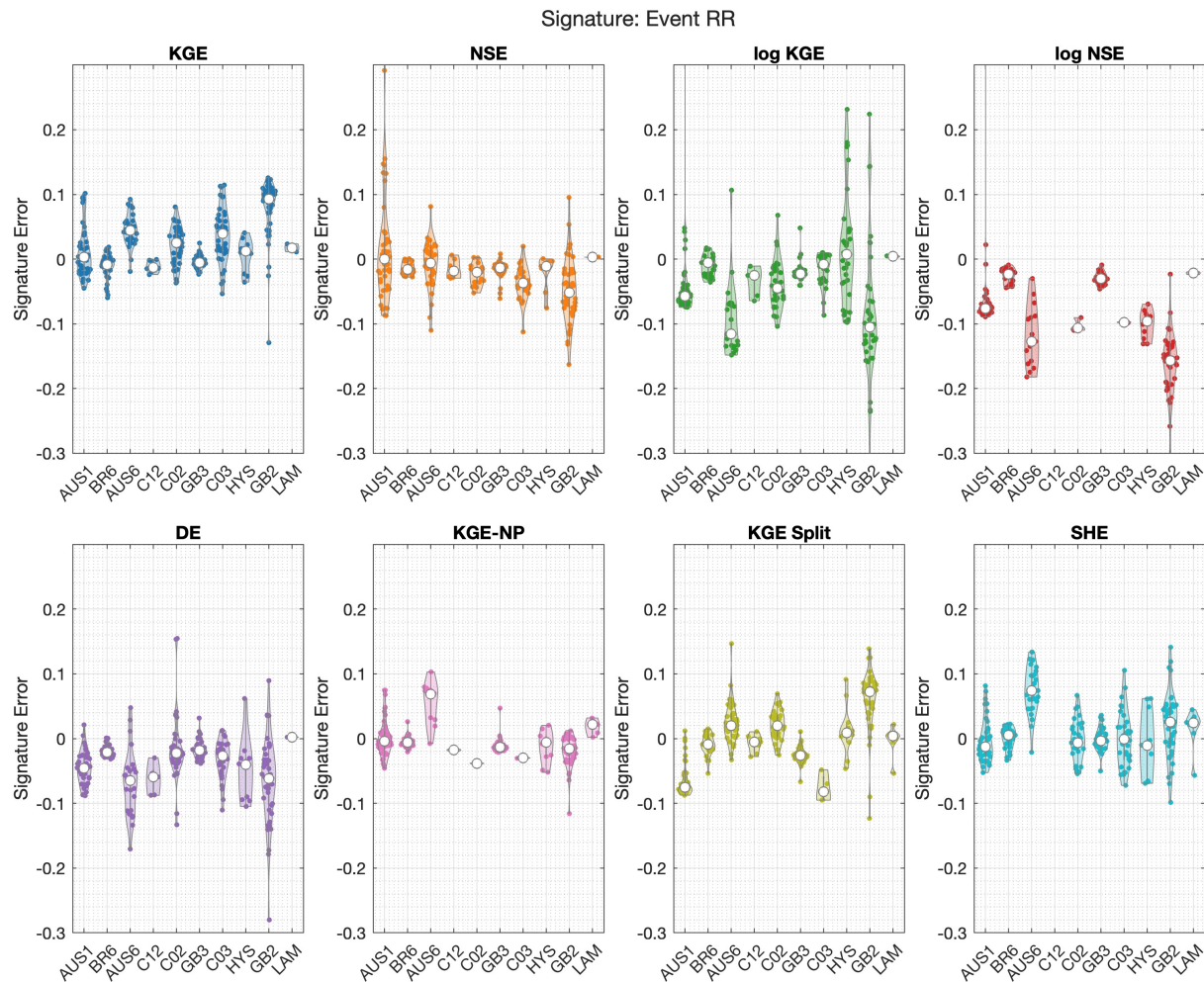


Figure S5. ERR for all OFs

Low Flow Percentile:

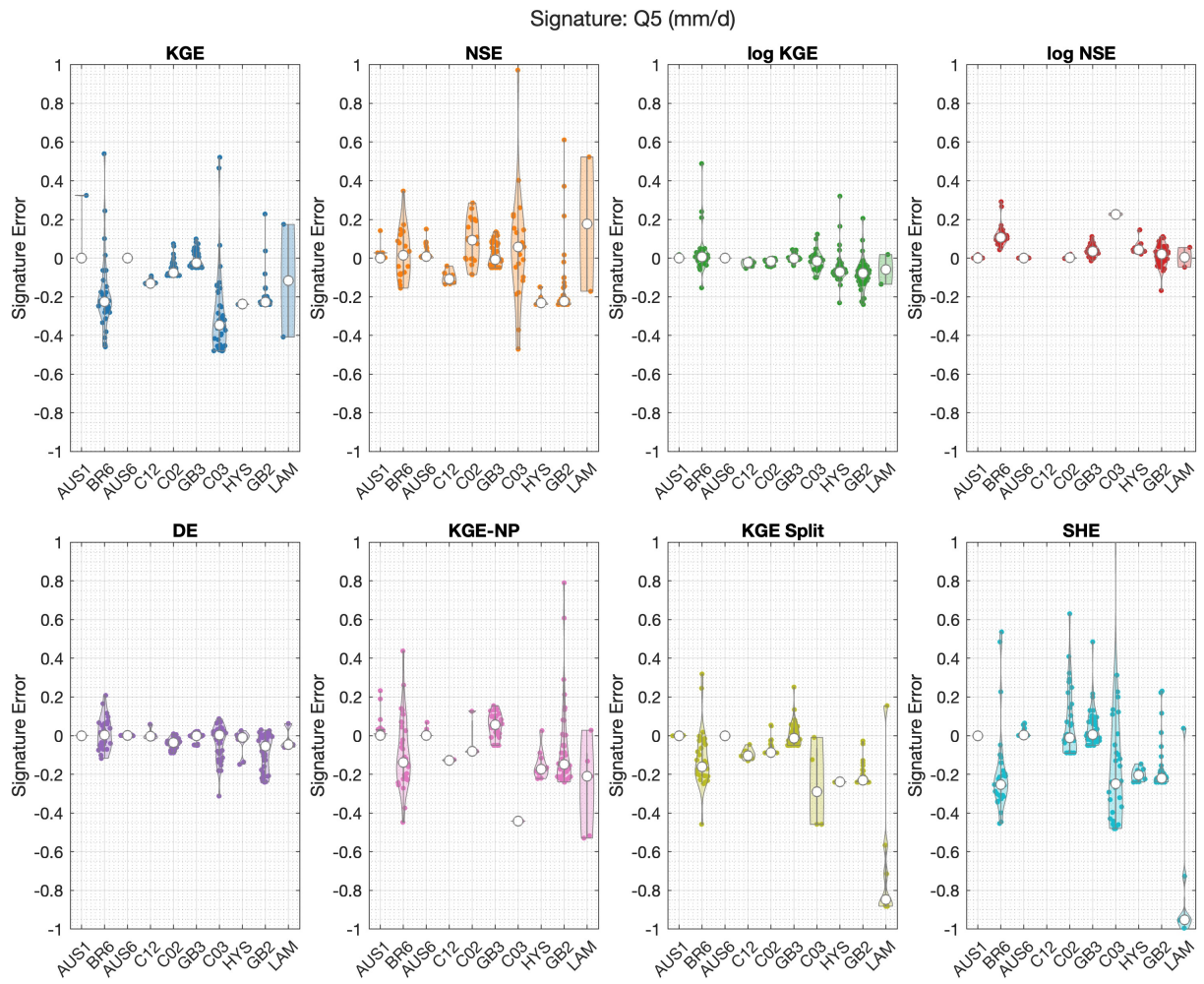


Figure S6. Details on Low Flow Percentile

High Flow Percentile:

Signature: Q95 (mm/d)

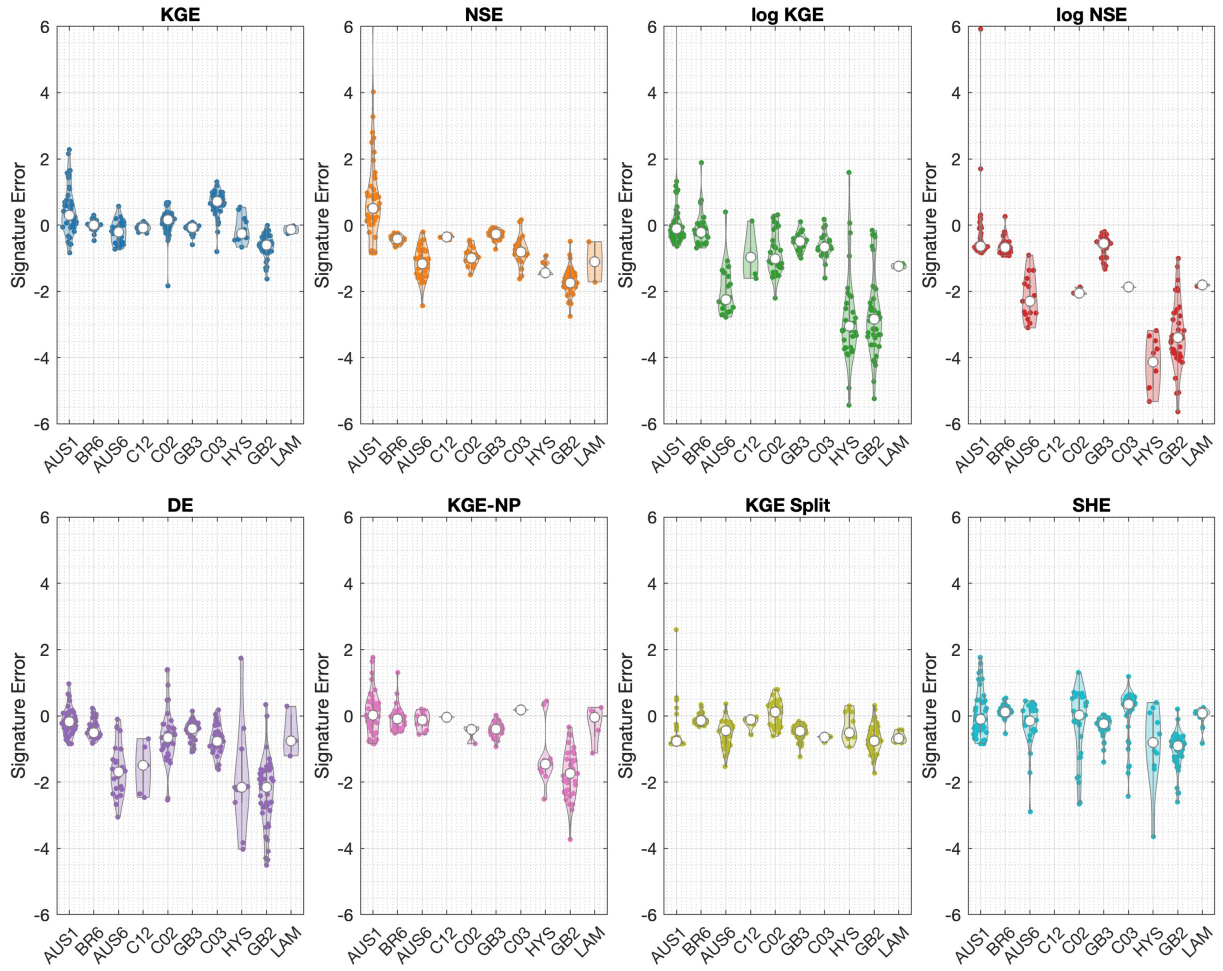


Figure S7. High Flow Percentile for all OFs

High Flow Duration:

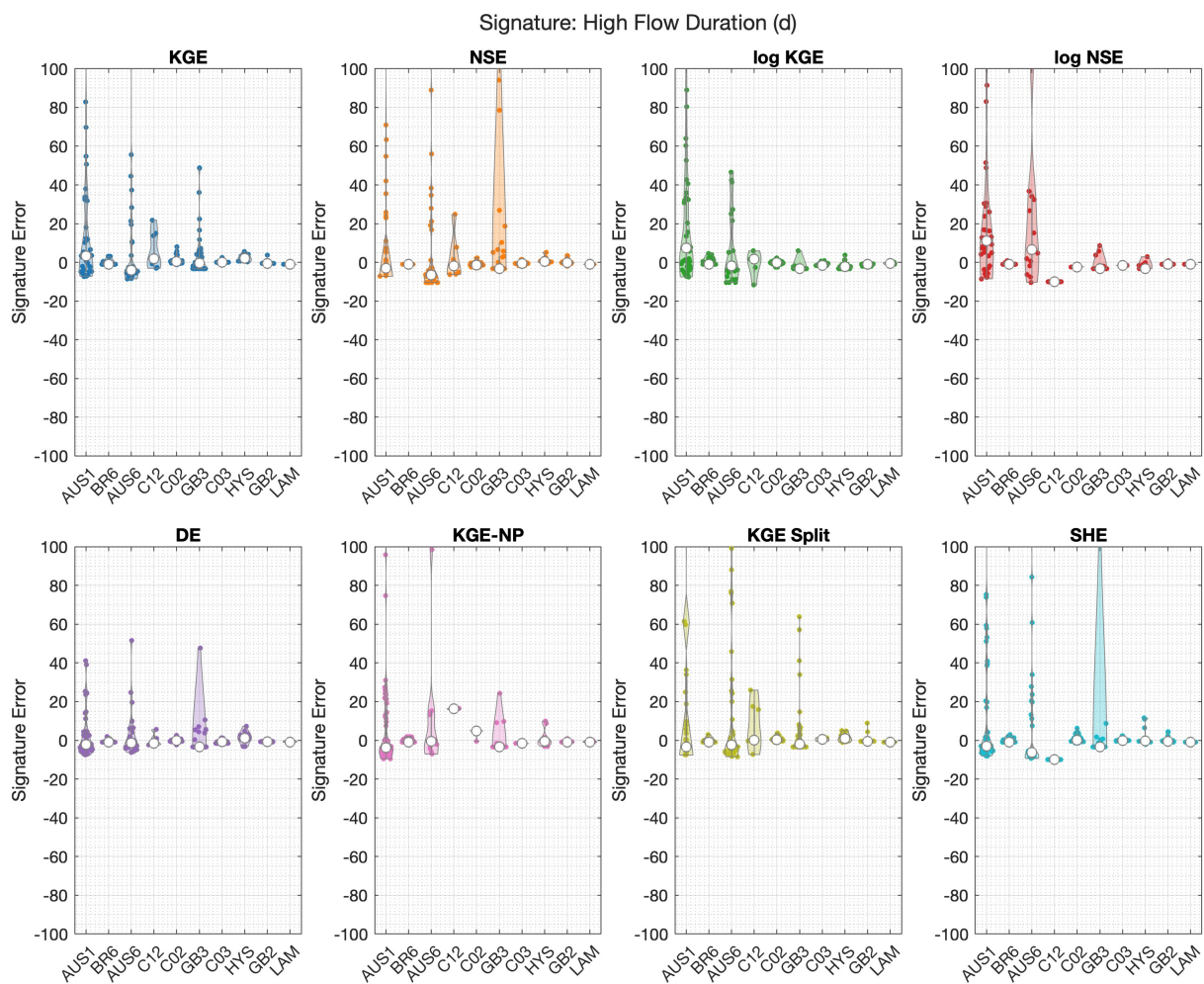


Figure S8. HFD for all OFs

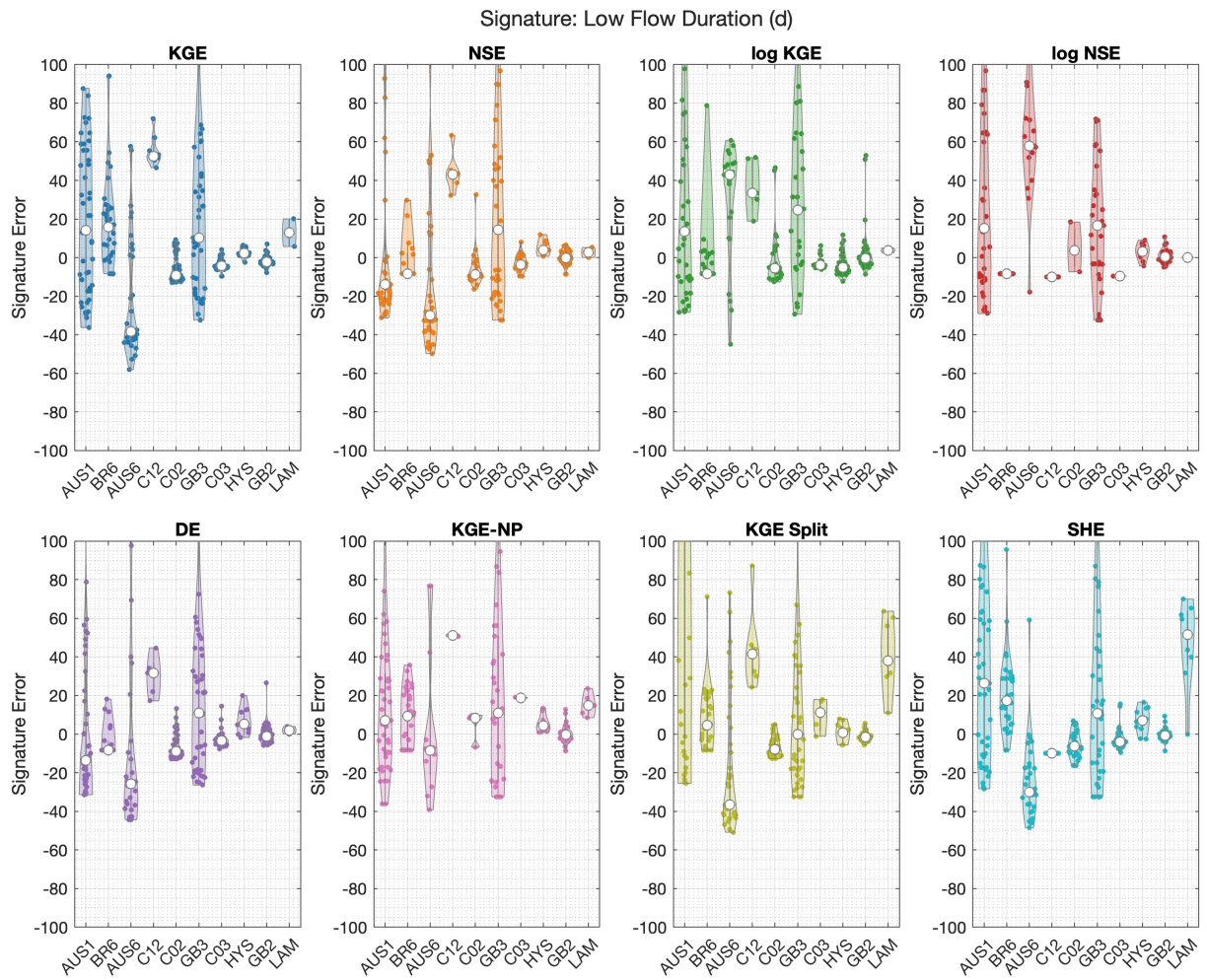


Figure S9. LFD for all OFs

High Flow Frequency:

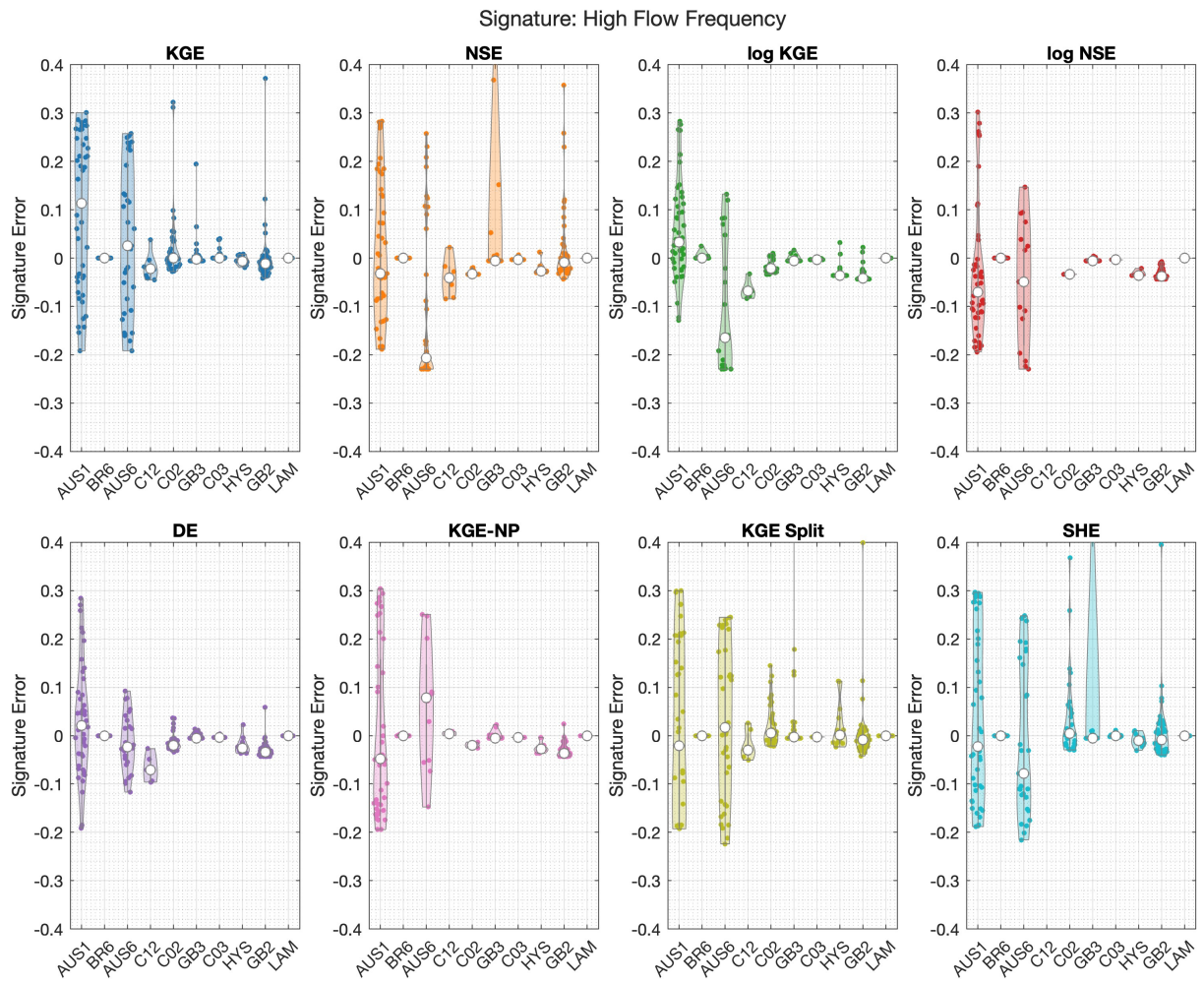


Figure S10. HFF for all OFs

Low Flow Frequency:

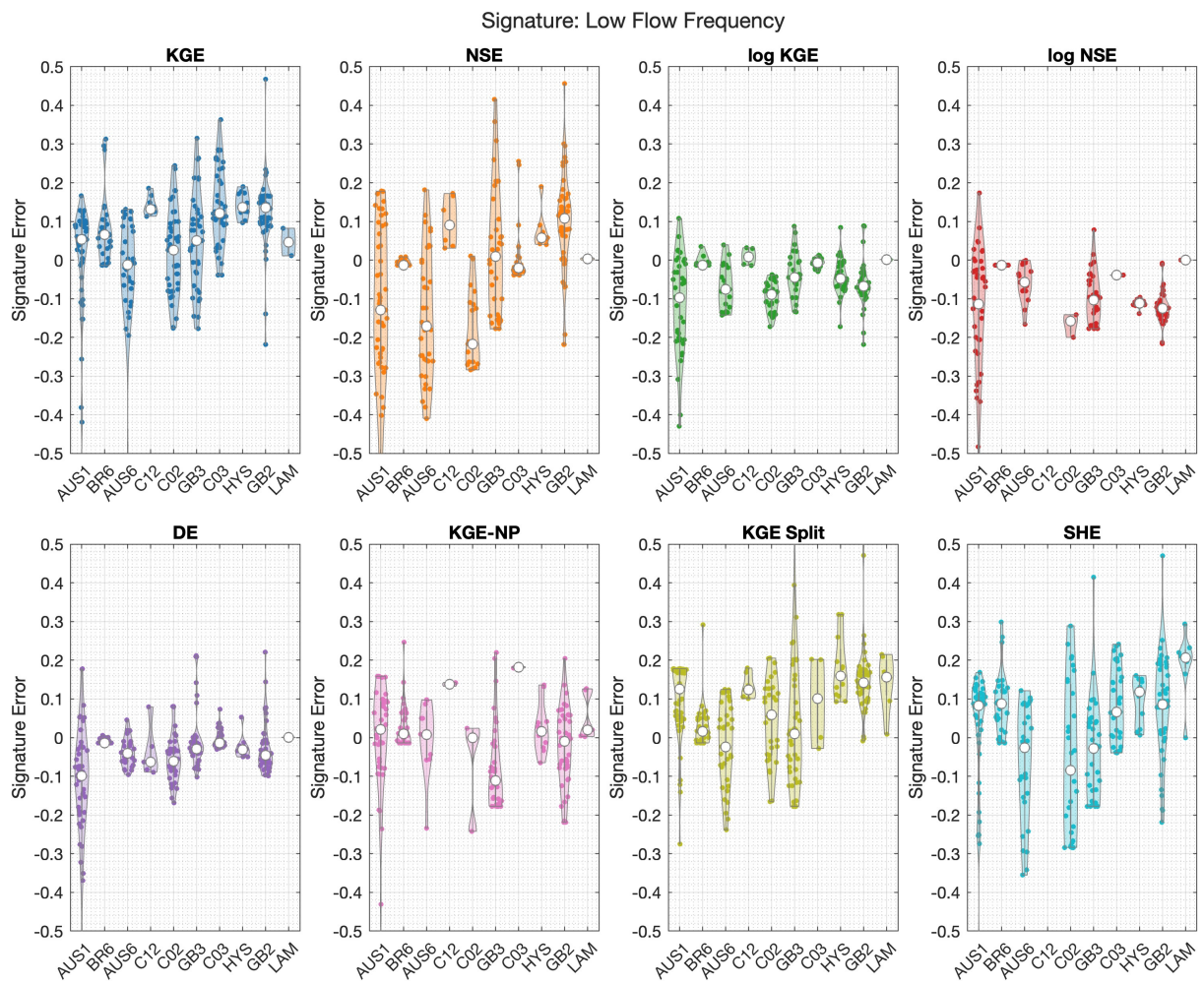


Figure S11. LFF for all OFs

Mean Half Flow Date:

Signature: Mean Half Flow Date (DOY)

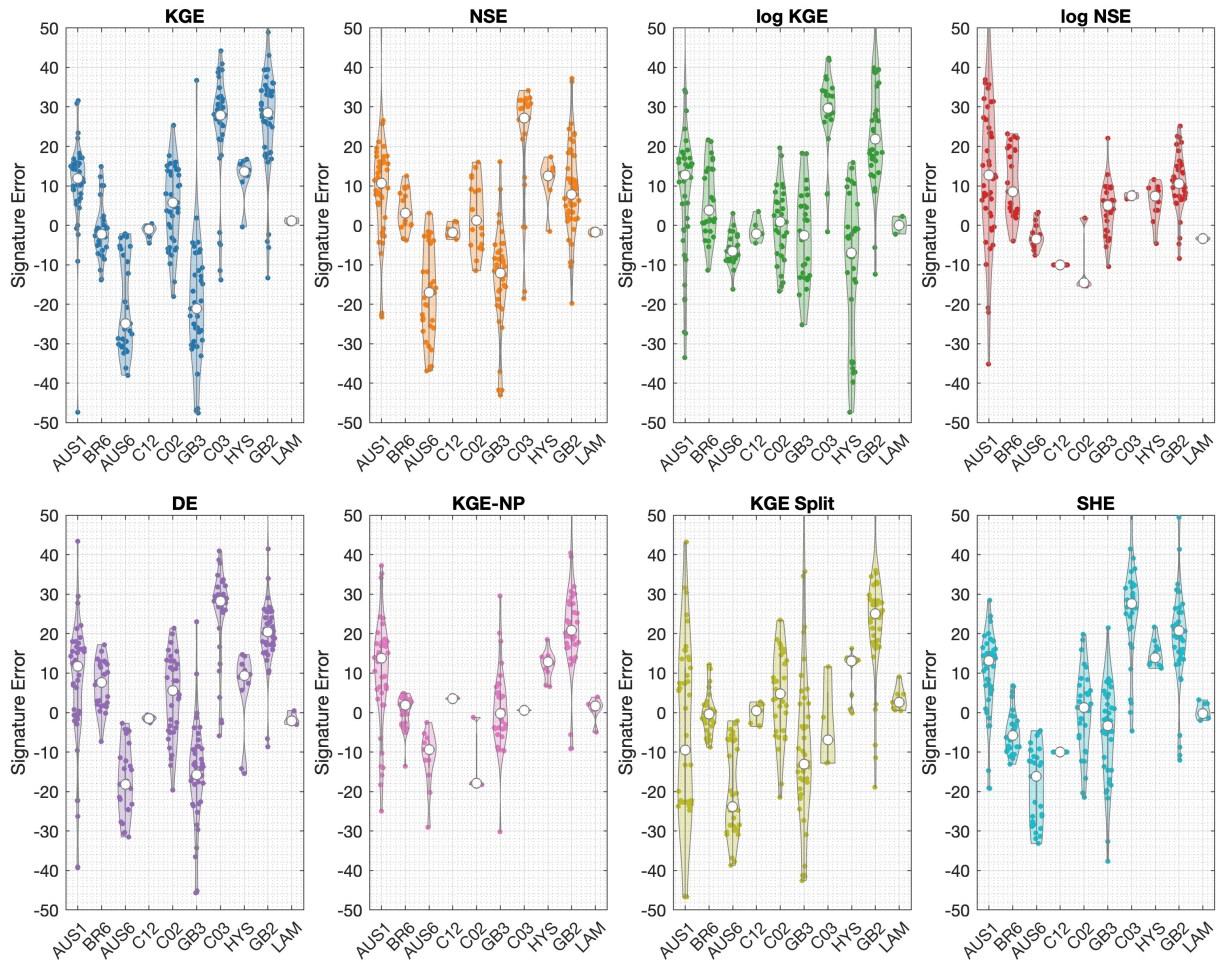


Figure S12. MHFD for all OFs

FDC Slope:

Signature: FDC Slope

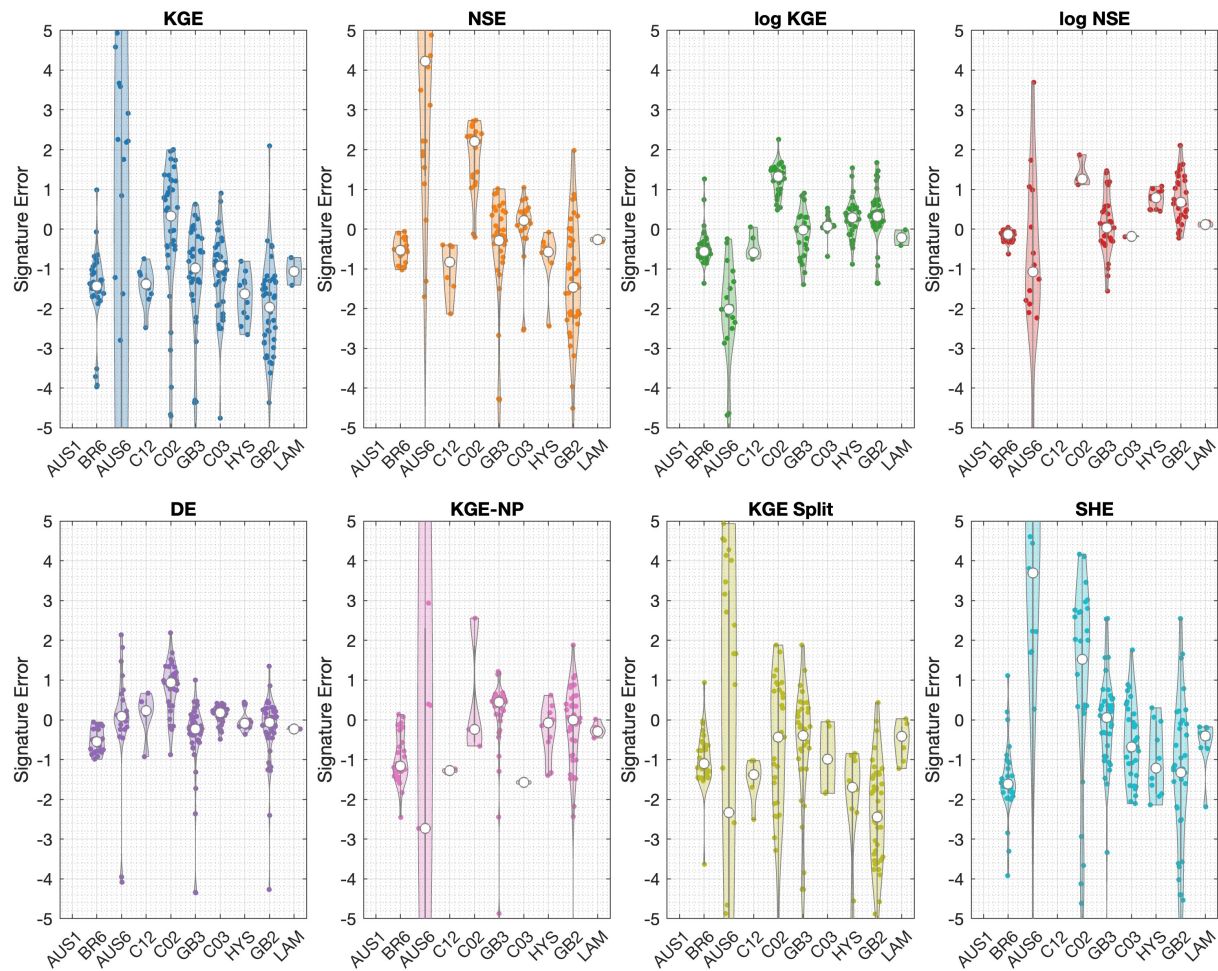


Figure S13. FDC Slope for all OFs

Variability Index:

Signature: Variability Index

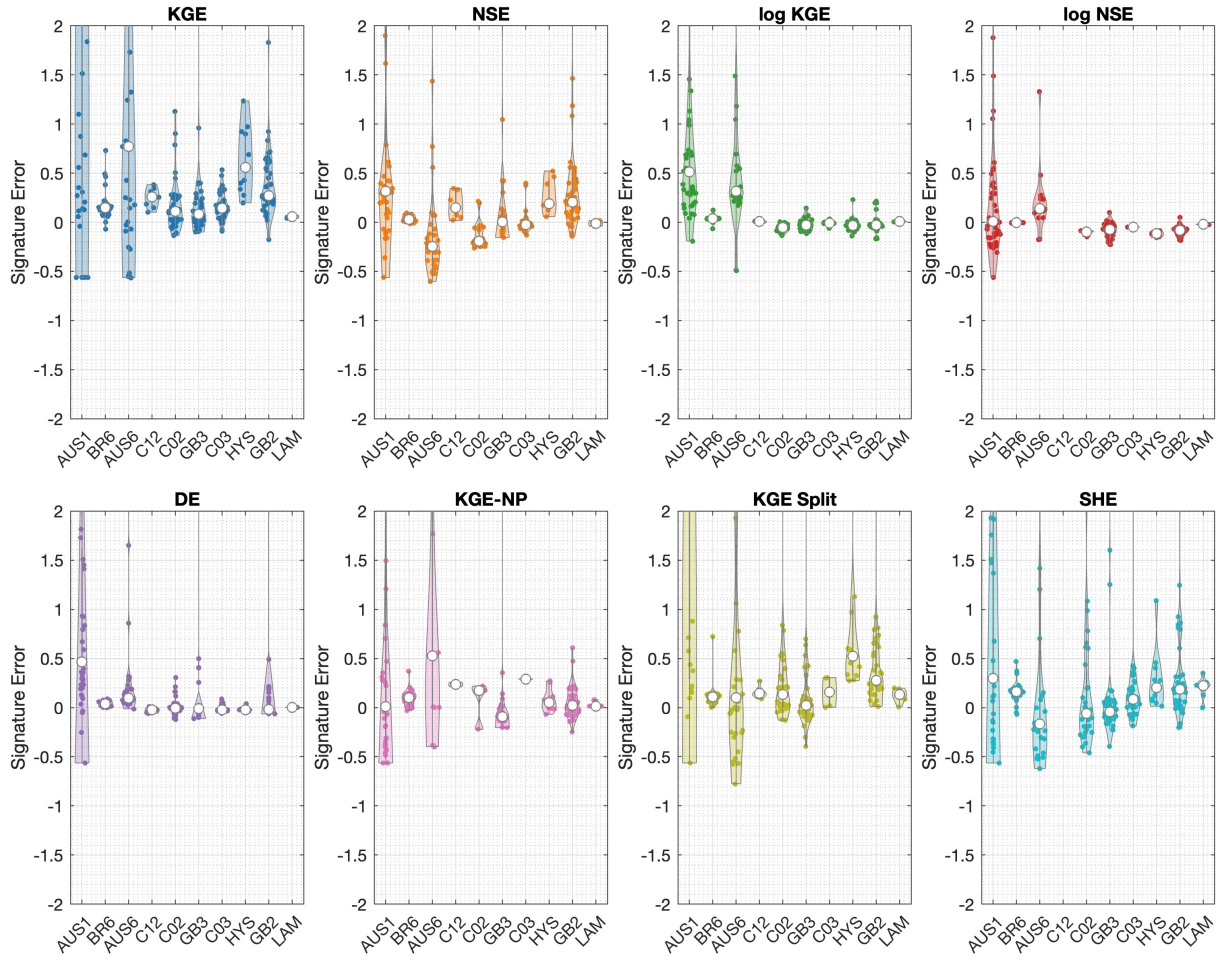


Figure S14. Variability Index for all OFs

Baseflow Index:

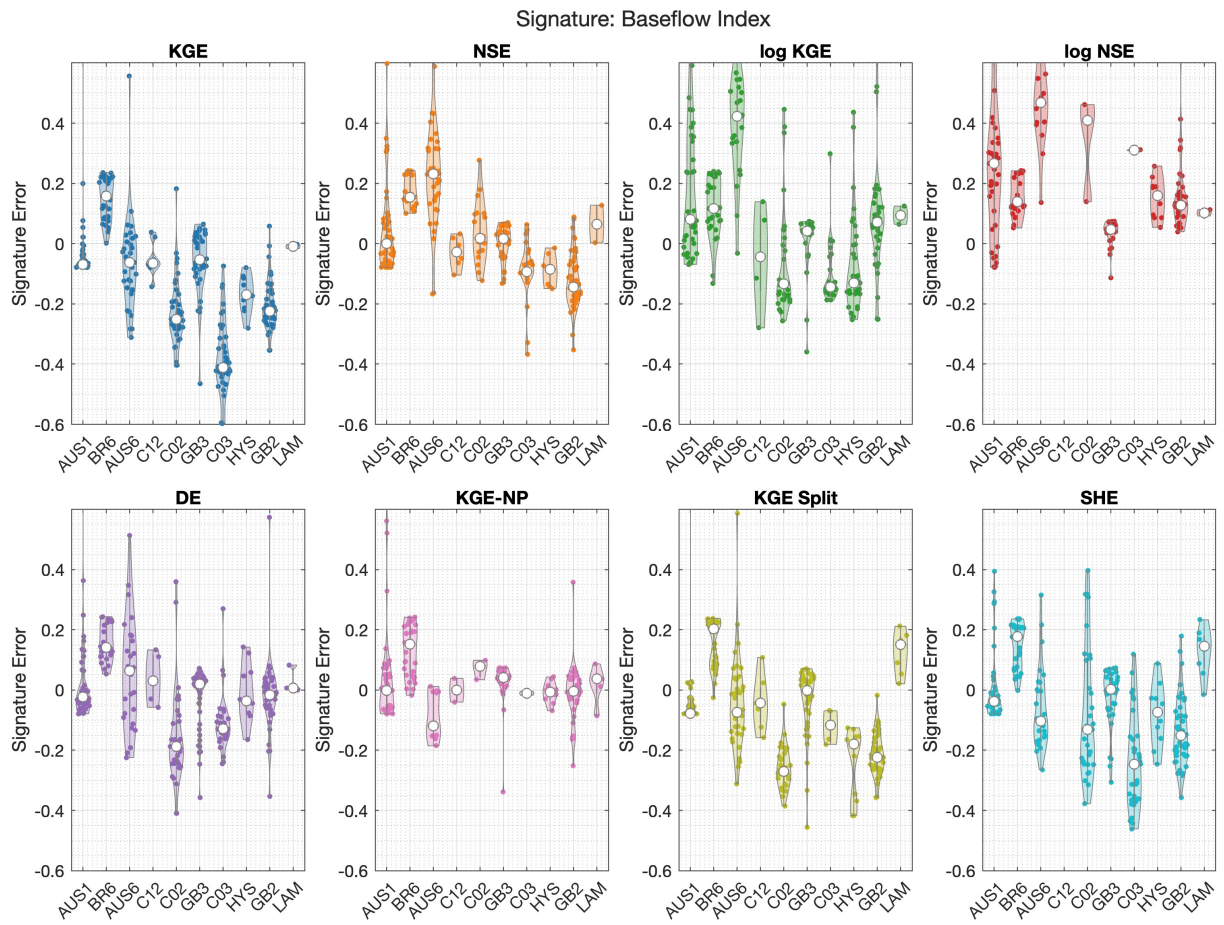


Figure S15. Baseflow Index for all OFs

Baseflow Recession Coefficient:

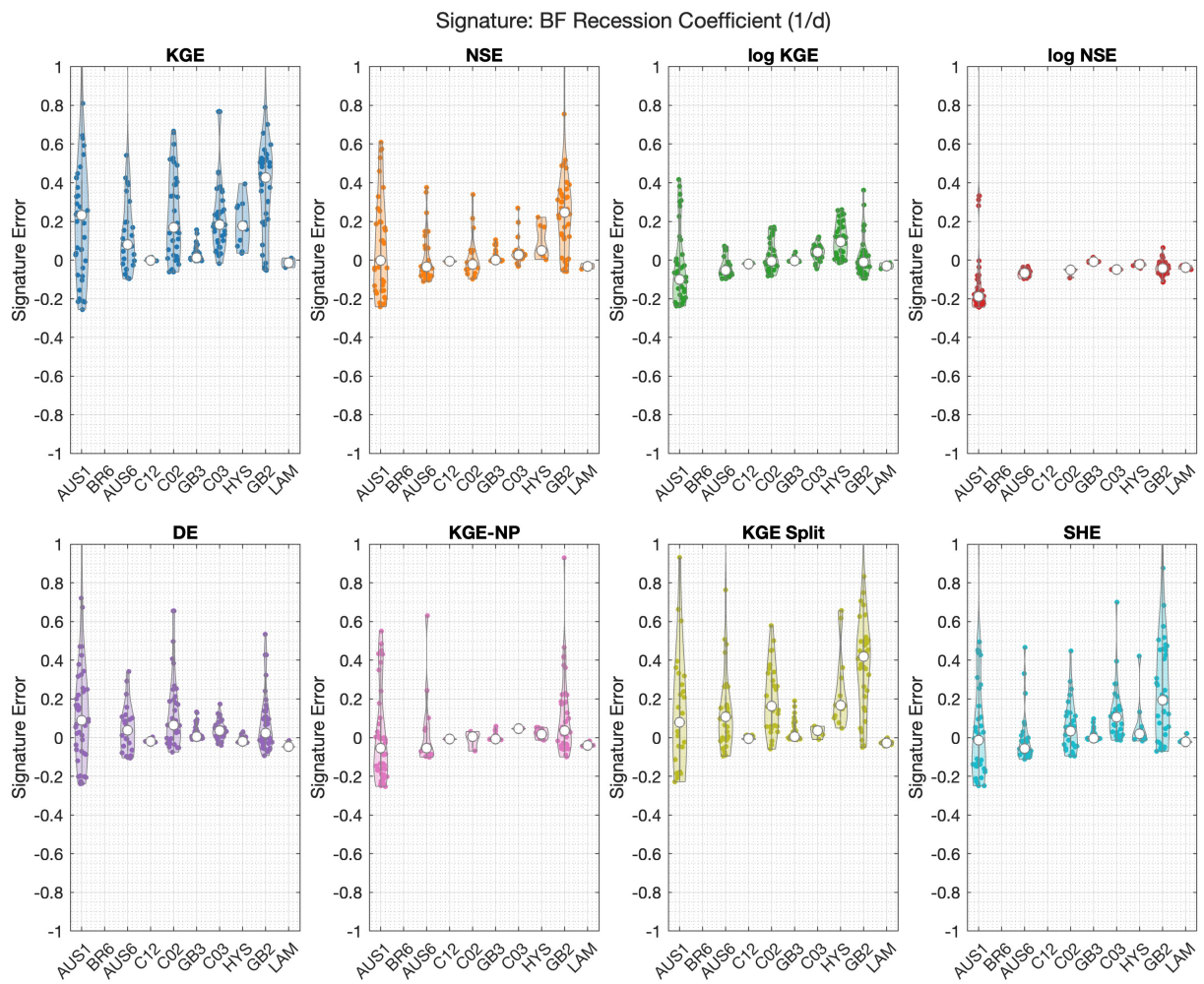


Figure S16. BRFC for all OFs

Flashiness Index:

Signature: Flashiness Index

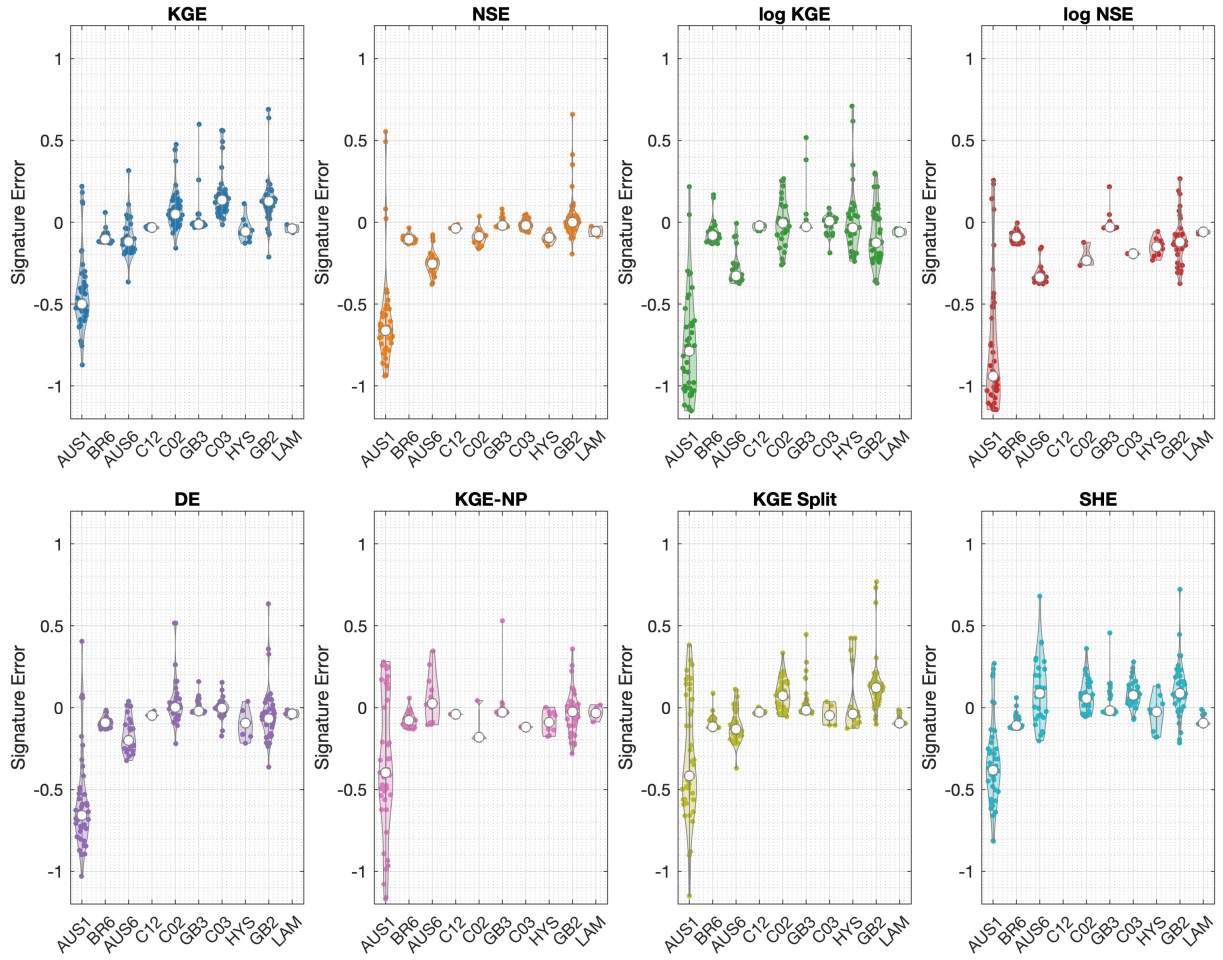


Figure S17. FI for all OFs

Rising Limb Density:

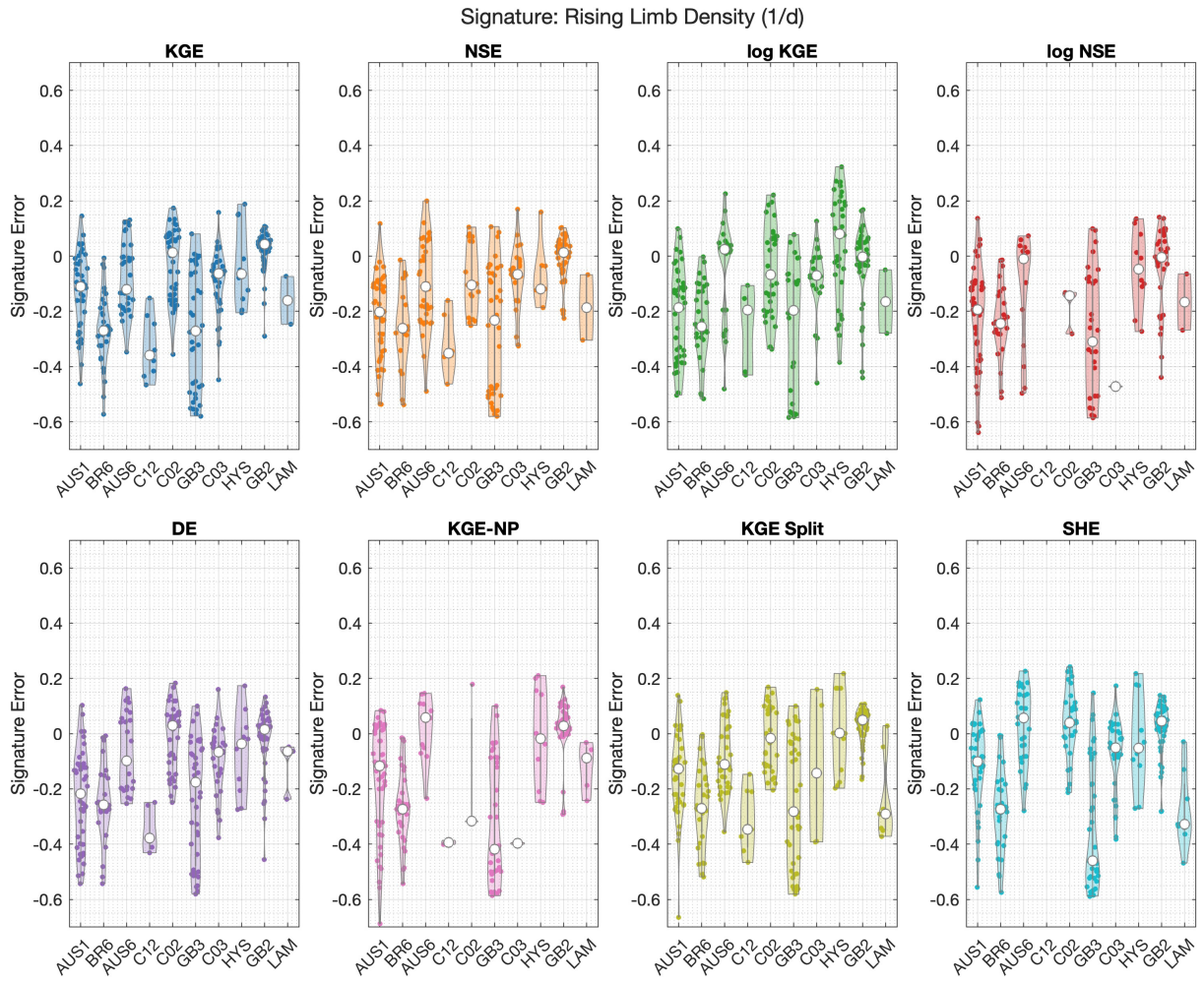


Figure S18. RLD for all OFs

Related to the BFI are also the Baseflow Recession Coefficient and the Flashiness Index. For these signatures, the case is less clear. For the BRFC, the least amount of uncertainty is seen for the log KGE and log NSE, making them the best OFs for this signature. This is also a second signature, where there is a trend such that the signature error has a positive trend relative to the humidity of the catchment.

For the Flashiness Index, it is hard to assess the best objective function. All OFs perform similarly in accuracy and variance, according to the quantitative analysis, KGE-Split, SHE, and KGE/KGE-NP are the best metrics to choose from. A clear effect is that models struggle to reproduce the flashiness of the ephemeral stream (AUS1). The vast majority of models, independent of calibration metrics, underestimate the flow. Nevertheless, there are models that are able to represent the FI signature values.

The last signature, that typically has significant influence from the OF choice is Event RR. Like the TRR, the best OF seems to be the KGE-NP, because it limits the range of the signature variation better than other OFs. Alternatives for the signature would be the NSE and the SHE.

S1.8 Table with Median t-values for statistical test

Table S3. Median p -values and fraction of tests with $p < 0.10$ and $p < 0.05$ across all objective function pairs for each hydrological signature.

Signature	Median p	$\#(p < 0.10)$ [%]	$\#(p < 0.05)$ [%]
Total RR (-)	0.029	71.4	57.1
Event RR (-)	0.045	64.3	50.0
Q95 (mm/d)	0.010	89.3	75.0
Q5 (mm/d)	0.087	53.6	46.4
BFI (-)	0.107	46.4	42.9
BFRC (-)	0.042	64.3	57.1
Flashiness Index (-)	0.075	53.6	39.3
LF Freq (-)	0.010	67.9	60.7
HF Freq (-)	0.301	14.3	7.1
FDC Slope (-)	0.191	35.7	14.3
Variability Index (-)	0.167	32.1	3.6
LF Dur (days)	0.292	17.9	7.1
HF Dur (days)	0.354	10.7	7.1
MHFD (DOY)	0.628	0.0	0.0
Rising Limb Density (-)	0.380	17.9	3.6

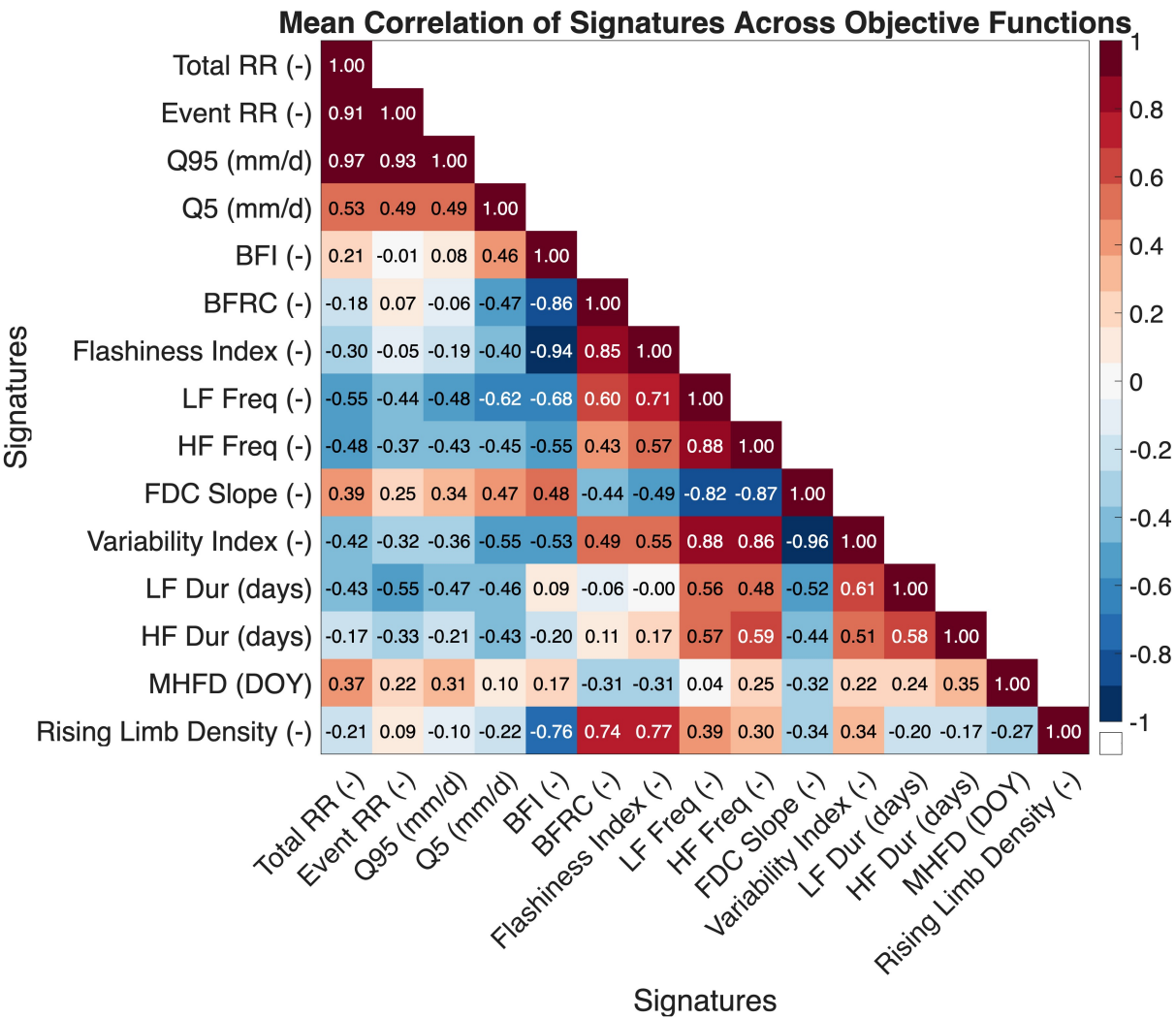


Figure S19. Signature Correlation Model-based

S1.10 Equifinality Assessment

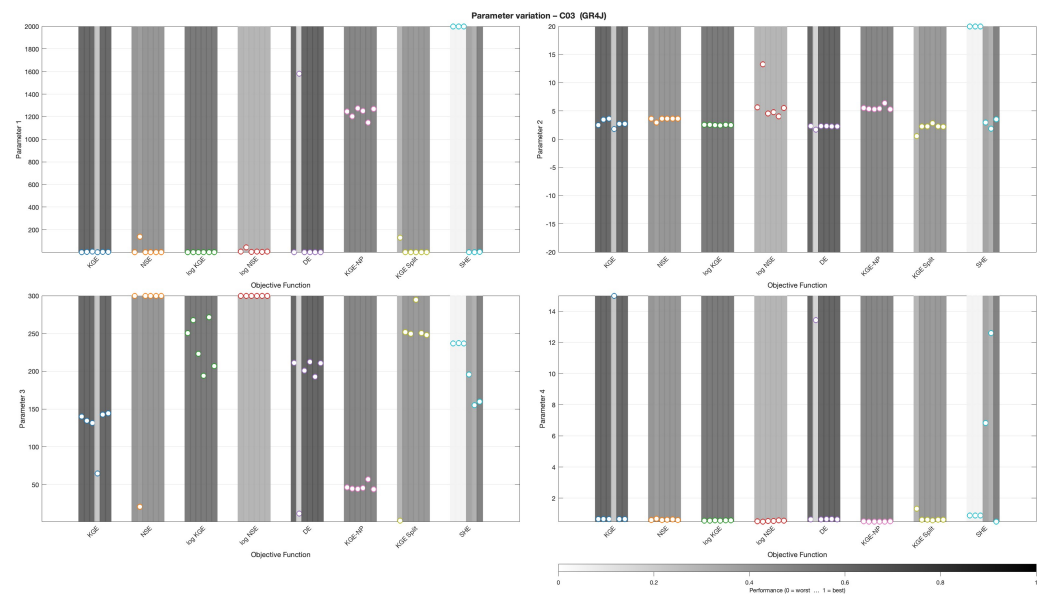


Figure S20. Assessment of Equifinality and Influence of Parameters. This plot shows the results exemplified for catchment C03 and GR4J in which many common patterns were found.

References

- Knoben, W. J. M., Woods, R. A., and Freer, J. E.: A Quantitative Hydrological Climate Classification Evaluated With Independent Streamflow Data, *Water Resources Research*, 54, 5088–5109, <https://doi.org/10.1029/2018WR022913>, 2018.
- Willmott, C. J. and Matsuura, K.: A More Rational Climatic Moisture Index*, *The Professional Geographer*, 44, 84–88, <https://doi.org/10.1111/j.0033-0124.1992.00084.x>, 1992.