UNIVERSIDAD DEL BÍO-BÍO

Interactive Discussion: Author Response to Reviewer 03

SPatial Efficiency And Kmoments (SPEAK): Evaluating Spatial Consistency in (Semi)Distributed Rainfall–Runoff Models

Matías Moreno, Pablo Mendoza, Eduardo Muñoz-Castro, Mauricio Zambrano-Bigiarini, and Alonso Pizarro
HESS discussions/ EGUsphere, doi: 10.5194/egusphere-2026-2912

RC: ***Reviewer Comment***, AR: Author Response, □ Manuscript text

Dear Reviewer 03 (Juraj Parajka),

Thank you for taking the time and effort to review this manuscript. Surely, incorporating your suggestions will improve the manuscript's readability and overall quality. Please find our responses to your comments below. These should be considered as preliminary (part of the interactive discussion) since the actual implementation of changes depends on the editorial decision. Below, we list reviewer's comments verbatim in bold, followed by responses to these comments in blue.

Thank you again for your helpful feedback!

Kind regards,

Alonso Pizarro (on behalf of all co-authors)

Sede Concepción
Avda. Collao 1202, Casilla 5-C - CP: 4051381. Fono/Fax: (56-41)3111200
Sede Chillán
Avda. Andrés Bello 720, Casilla 447 - CP: 3800708. Fono/Fax: (56-42)2463000
Email: ubb@ubiobio.cl
www.ubiobio.cl

RC 1: ***General comments: The manuscript proposes and evaluates a new spatial performance metric (SPEAK) for calibrating (semi)distributed rainfall–runoff models. The new framework is tested across 99 Chilean catchments, using ETa as a spatial calibration dataset. The presented objectives fit very well to the scope of HESS. The Introduction clearly presents the context, and the research gap (histogram sensitivity in SPAEF-type metrics) is reasonably well articulated. The proposed framework, which replaces histograms with K-moment-based smooth densities (KPDF) and a K-moment-based correlation (K-correlation) is a clear, well-motivated methodological contribution. The structure and presentation of the manuscript are generally clear, though some equations (rkm, KPDF derivation) lack an intuitive explanation for readers unfamiliar with the K-moment framework. The main weakness of the manuscript is that, in its current form, some interpretations would benefit from some additional analyses and supporting evidence related to independence and sensitivity of validation metrics, and the impact of the GLEAM resolution (and its refinement) on the evaluation of results.***

AR 1: We thank the reviewer for the positive assessment of the methodological motivation and for identifying the key evidential gaps.

RC 2: ***The applied "independent" validation metrics look not fully independent of the calibration objective. SPEAK and SPAEF both explicitly optimize a correlation-type term (rkm, r) and a CV-ratio term (βkm, β) during calibration. The paper then compares and validates using NRMSE and Pearson correlation, but Pearson correlation is structurally very similar to the SPAEF's own r term. Showing that SPEAK-calibrated fields have a higher Pearson correlation with the reference is not strong independent evidence, given that correlation-like objectives were already used directly during calibration. A more independent test would be to use metrics that were not part of either objective function, e.g., semivariogram/variogram range comparison, Moran's I (spatial autocorrelation structure), or a mutual information index between the simulated and reference fields. Moreover, the NRMSE improvements may be driven mostly by bias/magnitude correction, not spatial pattern. NRMSE is likely sensitive to the combined effects of bias and pattern errors. So, the interpretations will be better supported if the NRMSE is decomposed into bias and unbiased/pattern components.***

AR 2: Pearson correlation will no longer be described as an independent validation metric. To provide an independent spatial check, we will add one metric not used in SPEAK or SPAEF: the absolute difference between simulated and reference Moran's index. We will also report MSE, decomposed into mean-bias, standard-deviation, and correlation-related terms. This directly tests whether the NRMSE difference is driven by magnitude or by other reasons. See also AR 1 (anonymous reviewer 2).

RC 3: ***The interpretations from using GLEAM for validation have potentially the following limitations. Every 0.10° GLEAM native pixel becomes up to 4 correlated 0.05° pixels after interpolation. When SPEAK/SPAEF are computed "concatenated on time" across***

thousands of grid cells per catchment, a large fraction of those cells are not independent observations of spatial ETa variability. K-moment-based order statistics (Eq. 1) are computed over sample size n, and the true effective n is much smaller than the nominal n once interpolation redundancy is accounted for. This can affect interpretations that favor SPEAK, as it is unclear whether this reflects superior handling of hydrological heterogeneity or is a computational artifact. Another question concerns the fact that the number of catchments is smaller than that of a single native GLEAM 0.10° pixel. For such catchments, the reference dataset contains essentially no independent spatial information, yet these catchments are included in the summarized statistics. The manuscript/results/interpretations will benefit from stratifying results by catchment area relative to GLEAM's native pixel footprint (e.g., number of native 0.10° pixels per catchment), which can support/test whether SPEAK's advantage holds/increases if spatial information increases. A useful robustness check can also be if the SPEAK-vs-SPAEF comparison is repeated using GLEAM at its native 0.10° resolution (i.e. aggregating simulated ETa up to 0.10° rather than disaggregating GLEAM down to 0.05°).

AR 3: We agree and made three important methodological decisions: i) adopt the native 0.10° GLEAM spatial resolution for analysis (i.e., recalibrating all the catchments at that resolution), ii) apply a size filter to ensure at least 15 pixels at 0.10° within each catchment (See AR 5, Reviewer 1). The number of case studies was reduced from 99 to 43; and iii) include WSPAEF as a new objective function for calibration and comparison purposes.

We already recalibrated the TUW model at GLEAM's native spatial resolution (i.e., 0.10°). Three objective functions were considered (see also AR 1 from Anonymous Reviewer 2, and AR 3 from Reviewer 3 – Juraj Parajka): i) SPEAK; ii) SPAEF; and, iii) WSPAEF. Some results as follows:

a) Table R3.1 shows the 43 catchments with 15 or more pixels at a spatial resolution of 0.10°:

Table R3.1. Selected catchments for analysis at 0.10° spatial resolution (i.e., the native GLEAM resolution).

Catchment ID	N° of pixels at 0.10° spatial resolution
1310002	18
2101001	30
3414001	29
3430003	93
4314002	22
4513001	20
4523002	47
4703002	20
7112001	16
8117005	18

8123001	16
8124002	20
8132001	27
8135002	66
8317001	104
8334001	150
8342001	22
9123001	24
9129002	46
9135001	29
9140001	81
9404001	25
9414001	22
9420001	40
9437002	110
10111001	72
10134001	30
10304001	30
10328001	39
10340001	19
10356001	37
10364001	80
11141001	18
11143001	33
11302001	33
11308001	71
12284003	33
12284006	19
12284007	71
12582001	17
12802001	20
12805001	18
12876001	55

b) Reproduction of Figures 4, 5, and 6 at the new 0.10° spatial resolution:

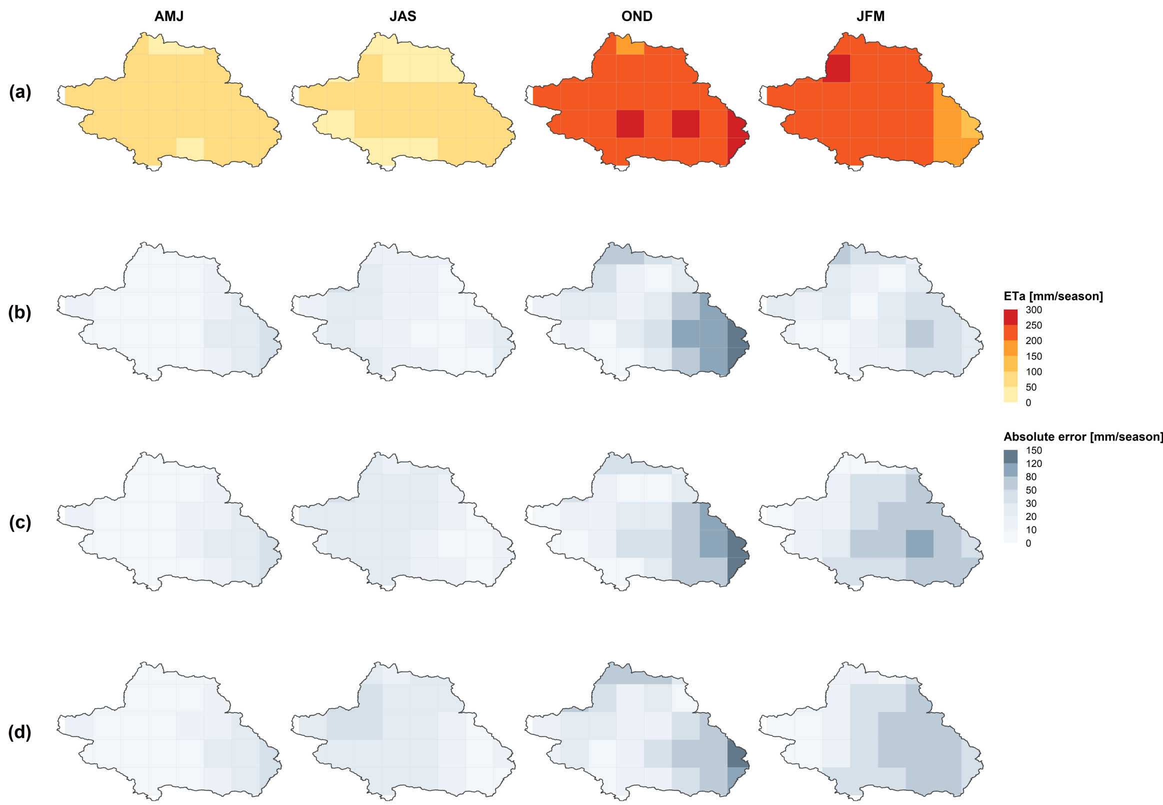

Figure R3.1 (Figure 4 in the main text): (a) Reference ETa from GLEAM for each three-month season (AMJ, JAS, OND, JFM), and spatial distribution of relative errors in simulated ETa with the TUW model with (b) the best SPEAK, (c) the best SPAEF, and (d) the best WSPAEF calibrations for catchment ID: 11302001 (period 2001-2020). The red colour bar indicates the ETa in (mm/season), whereas the grey colour bar indicates the absolute error (in mm/season) in simulated ETa with respect to the reference GLEAM data.

Figure R3.2 (Figure 5 in the main text): FSS for SPEAK-, SPAEF-, and WSPAEF-based mean annual ETa simulation (catchment ID: 11302001; period: 2001-2020). From a) to d): 75th, 90th, 95th, and 99th ETa percentiles taken as thresholds for FSS computation.

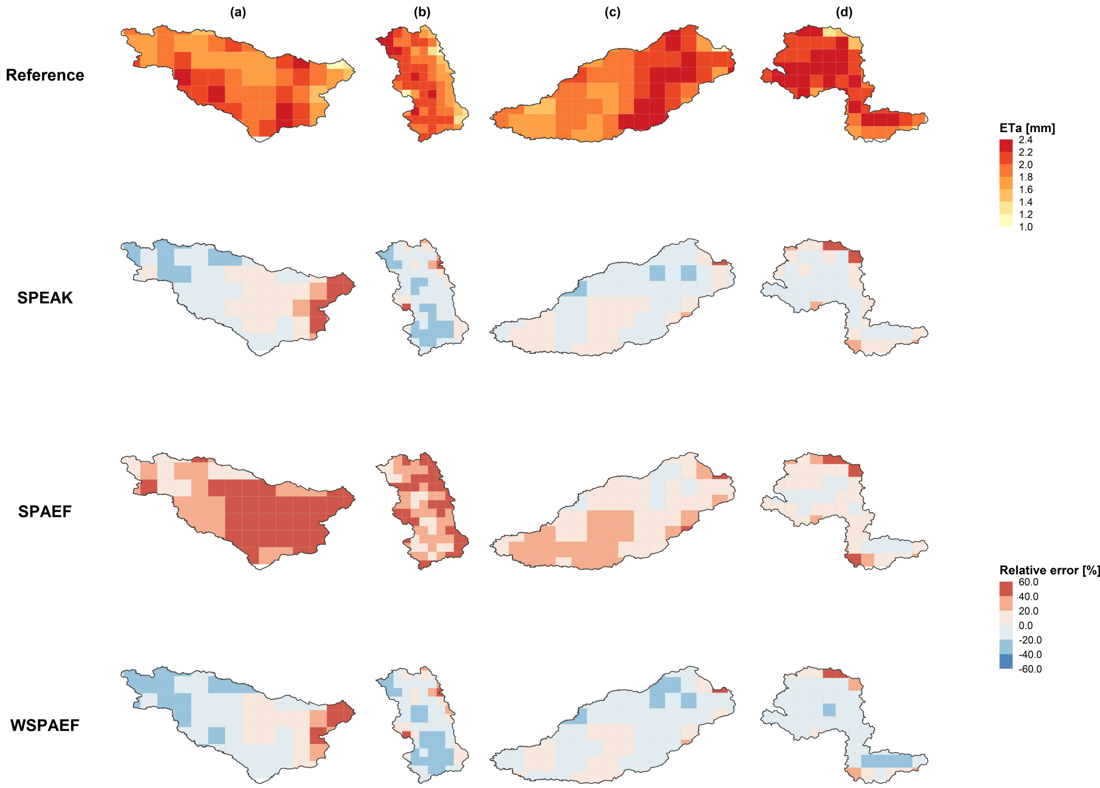

Figure R3.3 (Figure 6 in the main text): Comparison between GLEAM actual evapotranspiration fields (annual mean per pixel) and simulated ETa using parameters obtained by calibrating with SPEAK, SPAEF, and WSPAEF against GLEAM reference data. Each column corresponds to a specific catchment: (a) Río Itata at Balsa Nueva Aldea (ID: 8135002); (b) Río Biobío at Rucalhue (ID: 8317001); (c) Río Cautín at Almagro (ID: 9140001); and (d) Río Trancura Antes Río Llafenco (ID: 9414001).

c) Moran index and closer agreement with reference data:

Table R3.2. Moran index and closer agreement with GLEAM data.

Catchment ID	N° pixels at 0.10°	MORAN Index				Closer to GLEAM
		GLEAM	SPEAK	SPAEF	WSPAEF	
1310002	18	0.47	0.64	0.64	0.64	draw
2101001	30	0.42	0.77	0.57	0.77	SPAEF
3414001	29	0.35	0.39	0.40	0.69	SPEAK
3430003	93	0.70	0.87	0.91	0.87	SPEAK, WSPAEF
4314002	22	0.16	0.35	0.31	0.37	SPAEF
4513001	20	0.17	0.10	0.26	0.09	SPEAK
4523002	47	0.23	0.56	0.57	0.57	SPEAK
4703002	20	0.15	0.57	0.52	0.52	SPAEF, WSPAEF
7112001	16	0.64	0.14	0.83	0.09	SPAEF
8117005	18	0.32	0.83	0.75	0.86	SPAEF
8123001	16	0.38	0.36	0.26	0.42	SPEAK
8124002	20	0.31	0.61	0.40	0.62	SPAEF
8132001	27	0.48	0.61	0.35	0.70	SPEAK
8135002	66	0.50	0.87	0.75	0.89	SPAEF
8317001	104	0.63	0.53	0.22	0.53	SPEAK
8334001	150	0.69	0.66	0.60	0.72	SPEAK, WSPAEF
8342001	22	0.52	0.46	0.42	0.54	WSPAEF
9123001	24	0.31	0.25	0.08	0.19	SPEAK
9129002	46	0.39	0.75	0.65	0.73	SPAEF
9135001	29	0.58	0.67	0.61	0.67	SPAEF
9140001	81	0.52	0.82	0.79	0.84	SPAEF
9404001	25	0.20	0.08	-0.01	0.08	SPEAK, WSPAEF
9414001	22	0.31	0.20	0.21	0.30	WSPAEF
9420001	40	0.37	0.13	0.14	0.37	WSPAEF

9437002	110	0.43	0.85	0.84	0.86	**SPAEF**
10111001	72	0.31	0.56	0.57	0.61	**SPEAK**
10134001	30	0.06	0.55	0.34	0.51	**SPAEF**
10304001	30	0.38	0.36	0.25	0.29	**SPEAK**
10328001	39	0.44	0.77	0.84	0.79	**SPEAK**
10340001	19	0.17	0.04	0.41	0.09	**WSPAEF**
10356001	37	0.55	0.18	0.18	0.26	**WSPAEF**
10364001	80	0.61	0.84	0.87	0.86	**SPEAK**
11141001	18	0.37	0.50	0.65	0.67	**SPEAK**
11143001	33	0.51	0.85	0.83	0.86	**SPAEF**
11302001	33	0.29	0.85	0.86	0.85	**SPEAK, WSPAEF**
11308001	71	-0.03	0.89	0.90	0.92	**SPEAK**
12284003	33	0.26	0.48	0.65	0.47	**WSPAEF**
12284006	19	0.35	0.64	0.63	0.63	**SPAEF, WSPAEF**
12284007	71	0.51	0.78	-0.09	0.77	**WSPAEF**
12582001	17	0.07	0.61	0.70	0.72	**SPEAK**
12802001	20	0.27	0.66	0.67	0.66	**SPEAK, WSPAEF**
12805001	18	0.33	0.45	0.69	0.45	**SPEAK, WSPAEF**
12876001	55	0.53	0.55	0.82	0.55	**SPEAK, WSPAEF**
AVERAGE	**41.63**	**0.38**	**0.55**	**0.53**	**0.58**	
MIN	**16**	**-0.03**	**0.04**	**-0.09**	**0.08**	
MAX	**150**	**0.70**	**0.89**	**0.91**	**0.92**	

Sede Concepción
Avda. Collao 1202, Casilla 5-C - CP: 4051381. Fono/Fax: (56-41)3111200
Sede Chillán
Avda. Andrés Bello 720, Casilla 447 - CP: 3800708. Fono/Fax: (56-42)2463000
Email: ubb@ubiobio.cl
www.ubiobio.cl

d) Summary of results presented in AR 3.c:

Table R3.3. Hits (closer agreement with GLEAM data) as a function of SPEAK, SPAEF, and WSPAEF.

Objective function	Hits	Percentage
SPEAK	22	41%
SPAEF	15	28%
WSPAEF	17	31%

RC 4: Equations 1–8 are dense math with inconsistent notation between the main text and equations (e.g., sub/superscripts for K-moments are hard to follow without more explanation of what K′₁, K′₂ physically represent beyond Table 2).

AR 4: We agree and will rewrite the notation around a single glossary. Additionally, we performed numerical experiments to illustrate the advantage of using r_{km} and KPFD. We will prepare new supplementary material with the numerical experiments. See also AR 5.

RC 5: ***Eq. 4 (KCDF) is introduced, but its role in deriving Eq. 5 (KPDF) via "numerical differentiation" is not clear. Perhaps a worked example (even in supplementary material) showing how KPDF is computed from a sample would help readers unfamiliar with Koutsoyiannis's framework.***

AR 5: Thank you for pointing it out. Please, also see AR2 (anonymous reviewer 2).

Figures R3.4 and R3.5 show synthetic experiments to illustrate r_{km} and KDPF computations. These results will be documented in supplementary material, and code (in R) will be made available for reproducibility.

Figure R3.4 shows simulation results of standard correlation (r from Pearson correlation) and K-based correlation (r_{km}) for a pair of linearly dependent stochastic variables (Z and X, see below), obtained by 10 000 generated pairs. For the first variable (X), four different distributions were adopted (see panels clockwise): uniform, normal, lognormal with shape parameter $\varsigma = 1.00$, and Pareto with upper-tail index $\xi = 0.25$. The second variable is a weighed sum of the first one and an auxiliary variable with distribution identical to that of the first, where the weights are such that the correlation between the first and the second variable is a specified value ('theoretical r' in the graphs), i.e. $Z = rX + \sqrt{1 - r^2}\,Y$. Each plot has inside a subplot with a scatter plot for the particular theoretical case $r = 0.8$. We considered positive correlation values, as correlations are generally positive in hydroclimatic practice. Results provide evidence that r (Pearson) and r_{km} can be used interchangeably.

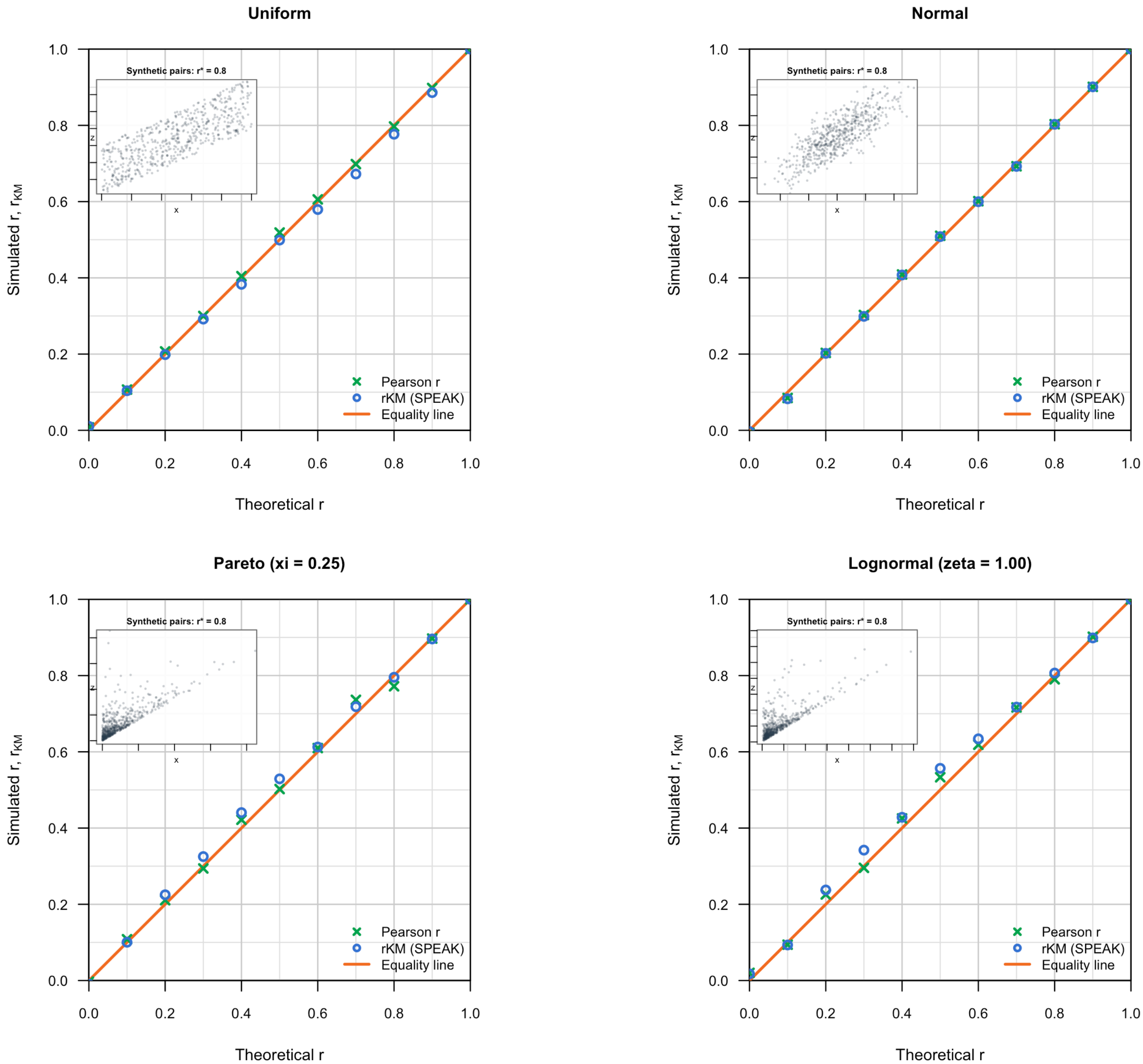

Figure R3.4. Synthetic experiments to compare standard correlation from Pearson (r) and K-based correlation (r_{KM}).

Additionally, Figure 3.5 shows a synthetic experiment in terms of probability density functions. For illustration purposes, a Pareto distribution with mean $\mu = 1$ and upper-tail index $\xi = 0.1$ was adopted. Left and right panels show Cartesian and logarithmic axes, respectively. Upper row figures show simulation results from a single data series of $n = 100$ values generated from the Pareto distribution. The points marked as "estimate" are computed with KPDF. For comparison, we also show a histogram. The lower panel shows simulation results from 100 data series of $n = 100$ values each, processed to calculate the median of the estimates and produce their uncertainty band in terms of prediction limits, where the original results are shown.

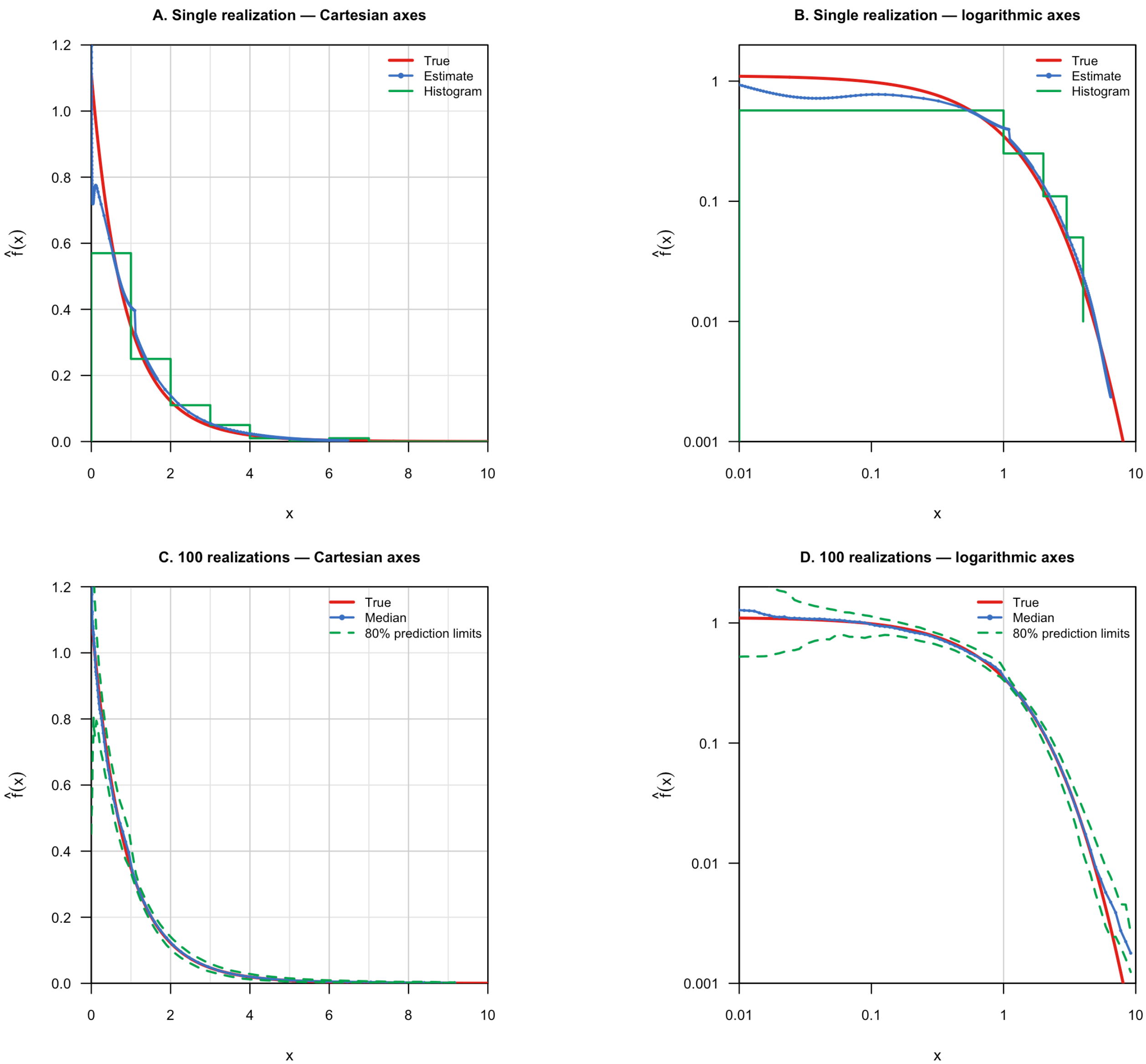

Figure R3.5. Synthetic experiments to illustrate KPDF estimation.

RC 6: ***The rkm formula (Eq. 3) is presented without much intuition — it is not obvious why a "max ratio" formulation produces something behaving like a correlation coefficient bounded appropriately; a brief justification or reference to boundedness proofs would help.***

AR 6: Please, see AR 5. More information (including explicit rationale) about r_{km} and KDPF will be available in the new supplementary material. Related codes (in R) will also be available for reproducibility.

RC 7: ***The discussion connects reasonably well to previous literature on spatial calibration, however the manuscript will benefit from some additional comparisons of SPEAK against the other SPAEF variants listed in Table 1. Given Table 1 provides alternative competing metrics, so benchmarking against 1–2 other alternatives (e.g., WSPAEF, which also addresses histogram sensitivity) would substantially strengthen/support the novelty claims.***

AR 7: We agree that WSPAEF is the most relevant additional comparator because it also addresses histogram sensitivity. Calibration with WSPAEF was performed during the review period. See AR3 (Reviewer 3) and AR14 (Reviewer 2) for some results.

RC 8: ***The results and interpretations do not include calibration uncertainty/equifinality assessment. The results (especially catchment-count comparisons like "85 of 99 catchments") could be sensitive to the calibration method. Please consider discussing whether the SPEAK/SPAEF differences reflect more the metric difference rather than the variability in the calibration approach.***

AR 8: We thank the reviewer for raising this important point. We agree that calibration uncertainty and parameter equifinality were not explicitly quantified in the manuscript as was submitted. Fortunately, hydroPSO (the R package used for calibration) has a built-in module for uncertainty analysis. Results will be available in the new manuscript version.