UNIVERSIDAD DEL BÍO-BÍO

Interactive Discussion: Author Response to Reviewer 02

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 02 (Anonymous Referee),

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: ***The manuscript by Moreno et al. presents a variation of an existing metric used for hydrological model calibration. The proposed metric is intended to improve the spatial calibration of evapotranspiration (ET) patterns in distributed hydrological models.***

While I agree with the authors that there is room for improving existing spatial performance metrics, I am not convinced that the presented modelling experiments are sufficient to demonstrate the advantages and general applicability of the proposed approach. Furthermore, I find that several of the presented results focus on grid-based ET errors, although the spatial performance metrics investigated (SPAEF and SPEAK) are insensitive to biases. Consequently, some of the presented evaluations appear to be of limited relevance.

I am also not convinced that the selected hydrological model setup/code is appropriate for developing and testing the proposed metric. Although the model is described as semi-distributed or fully distributed, only precipitation and air temperature are spatially distributed forcing variables. Most of the remaining catchment characteristics appear to be spatially uniform because they are derived from the catchment-aggregated CAMELS-CL dataset. Ideally, a study aiming to evaluate spatial ET performance metrics should employ a model with spatially distributed parameters and process representations that are sensitive to ET variability, such as spatially distributed soils, vegetation, and land use.

In addition, the spatial resolution of GLEAM is relatively coarse (approximately 100 km² per grid cell). Many of the investigated catchments contain only around 20 GLEAM cells (most catchments are substantially smaller than 2,000 km²). This considerably limits the amount of spatial information available for evaluating spatial ET patterns.

Overall, while the manuscript addresses an interesting topic, I do not believe that the presented experiments provide sufficient evidence to support the claimed benefits of the proposed metric. Therefore, I cannot recommend the manuscript for publication in HESS in its current form.

AR 1: Thank you for this comprehensive assessment. We agree that the original manuscript did not adequately distinguish between magnitude and spatial-pattern agreement, did not sufficiently account for the effective spatial resolution of the native GLEAM product, and stated conclusions that extended beyond the modelling configuration evaluated. We have therefore revised the evaluation framework and narrowed the scope of our conclusions. We address the reviewer's main concerns below:

- **Modelling experiments:**

  While we agree that SPAEF was intentionally formulated to be insensitive to positive multiplicative bias, SPEAK, however, is not completely bias-insensitive. For instance, if the simulated field is a positive rescaling of the reference field,

$$S = aR, \qquad a > 0,$$

Pearson correlation remains equal to one, the coefficient of variation remains unchanged, and the z-score distributions used in the histogram-intersection component are identical. Consequently, SPAEF can equal one even when $a \neq 1$. SPAEF therefore evaluates similarity in spatial structure while largely disregarding multiplicative differences in magnitude.

SPEAK, however, is not completely bias-insensitive. Its r_{KM} component and K-moment-based coefficient-of-variation ratio are invariant to positive multiplicative scaling. In contrast, the γ_{KM} component is defined as

$$\gamma_{KM} = 1 - H(P_R, P_S),$$

where $H(P_R, P_S)$ is the Hellinger distance between the reference and simulated KPDFs, and R and S mean reference and simulated variable. In the present implementation, we construct these KPDFs from the original, non-standardised ETa values and evaluate them over a common domain. A change in distributional location or scale therefore produces $H(P_R, P_S) > 0$ and $\gamma_{KM} < 1$. SPEAK is consequently only partially scale-insensitive and, unlike SPAEF, can respond to magnitude differences through its KPDF component.

Nevertheless, we agree that NRMSE alone is not sufficient evidence of improved spatial-pattern reproduction because it combines systematic magnitude error and centred pattern error. We will therefore decompose the mean-square error as

$$\mathrm{MSE} = \left(\overline{S} - \overline{R}\right)^2 + (\sigma_S^2 + \sigma_R^2 - 2r\sigma_S\sigma_R),$$

where the first term represents squared mean bias and the second represents centred error associated with variability and spatial correspondence. The revised results will report both components separately. We will interpret NRMSE as an overall magnitude-sensitive error measure rather than as a pure spatial-pattern metric. We will also include, by catchment, the absolute difference in Moran's index (I), $|I_S - I_R|$, as an independent diagnostic of spatial autocorrelation that is not explicitly included in either calibration objective.

- **Lumped and (semi) distributed experiments:**

Second, we decided to use spatially uniform parameter sets to simplify the calibration experiments, obtaining spatial variability in model states and fluxes only due to spatial heterogeneities in meteorological fields. In the revised manuscript, we will correct the model description, rename the 'fully distributed' case, and remove process claims involving spatial soil, vegetation, or land-use controls. The following sentences will be added to the main text:

a) Fully distributed model configuration will be renamed as 'gridded-forcing configuration':

… In the gridded-forcing configuration, precipitation, temperature, and PET were supplied on the 0.10° spatial resolution (i.e., native GLEAM resolution), while one calibrated TUW parameter vector was applied uniformly to all grid cells within each catchment. Thus, this configuration represents spatial variability induced by the meteorological forcings and model states, but not spatial heterogeneity in soil-, vegetation-, or land-use-dependent parameters. Testing SPEAK in a model with pedotransfer functions or parameter regionalisation remains an important next step.…

b) Discussion section with related information and new research paths:

… Future research will attempt to test the potential of SPEAK using a spatial parameter configuration that considers pedotransfer functions or other spatial regularization strategies.…

- **Adopted spatial resolution for analysis:**

Third, we agree that bilinearly interpolating GLEAM from 0.10° to 0.05° does not create additional independent spatial information. As a consequence, all catchments were recalibrated at 0.10° (native GLEAM resolution). See results in AR6 and AR14.

Finally, additions or modifications (preliminary) to the main text as follows:

… SPAEF is invariant to positive multiplicative rescaling because Pearson correlation, the coefficient of variation, and the histogram overlap between z-standardised values remain unchanged when $S = aR$, with $a > 0$. SPEAK is only partially scale-insensitive. Although its r_{KM} and β_{KM} components are invariant to positive multiplicative scaling, its γ_{KM} component compares KPDFs constructed from the original, non-standardised values and therefore responds to differences in distributional location and scale…

… Because NRMSE combines systematic magnitude differences and centred pattern errors, the mean-square error was decomposed as $\mathrm{MSE} = \left(\overline{S} - \overline{R}\right)^2 + (\sigma_S^2 + \sigma_R^2 - 2r\sigma_S\sigma_R)$. The first component represents squared mean bias, whereas the second represents centred error associated with variability and spatial correspondence. Differences in Moran's index (I) between simulated and reference ETa fields were additionally used to evaluate spatial autocorrelation independently of the calibration objectives…

… To avoid interpolation-induced issues, CR2MET forcing are aggregated from 0.05° to 0.10° to match the native 0.10° GLEAM grid resolution…

…This study evaluates SPEAK using one conceptual hydrological model under semi-distributed and gridded-forcing configurations with spatially uniform calibrated parameter values within each catchment and using one ETa product. The results therefore characterise the behaviour of SPEAK within this controlled experimental setting but do not establish universal superiority across hydrological models, spatial parameterisations, or ET products…

RC 2: ***Section 2.1: The introduction of this section is unnecessarily broad. The manuscript ultimately evaluates spatial patterns of long-term averaged ET, and therefore statements such as: "This phenomenon is particularly pronounced in hydrology, where observations are typically limited to a few decades, whereas the processes of interest may operate on centennial or millennial time scales." do not appear directly relevant to the scope of this study.***

AR 2: Thank you for this comment. We agree that the quoted statement concerning centennial or millennial hydrological timescales is not directly required to justify the application presented here. As a consequence, the sentence in question will be removed. However, we do not consider the methodological rationale for introducing K-moments to be unnecessary. Section 2.1 is intended to explain why order-statistic-based descriptors are introduced before presenting r_{KM}, β_{KM}, and the KPDF. This rationale is directly connected to the formulation of SPEAK.

It is important to note that SPEAK and SPAEF are not calibrated exclusively against long-term averaged ETa patterns. Instead, both objective functions are computed from daily reference and simulated ETa maps concatenated over the calibration period. Temporally aggregated seasonal or long-term maps are subsequently used for presentation and diagnostic evaluation. Therefore, the empirical samples underlying the calibration retain temporal as well as spatial variability, although we agree that this does not make the discussion of centennial or millennial processes relevant.

RC 3: ***Section 2.2: To better illustrate the advantages of the proposed metric (rkm) over conventional correlation measures, the manuscript would benefit from simple synthetic examples presented as scatter plots. Likewise, the KPDF component could be explained using synthetic examples. Such illustrations would make the methodology considerably more accessible to the broader hydrological community.***

AR 3: We agree. We provide, as follows, results of synthetic experiments that will be available in in the new supplementary material (see also AR 5 and AR 6 from Reviewer 3):

Figure R2.1 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.

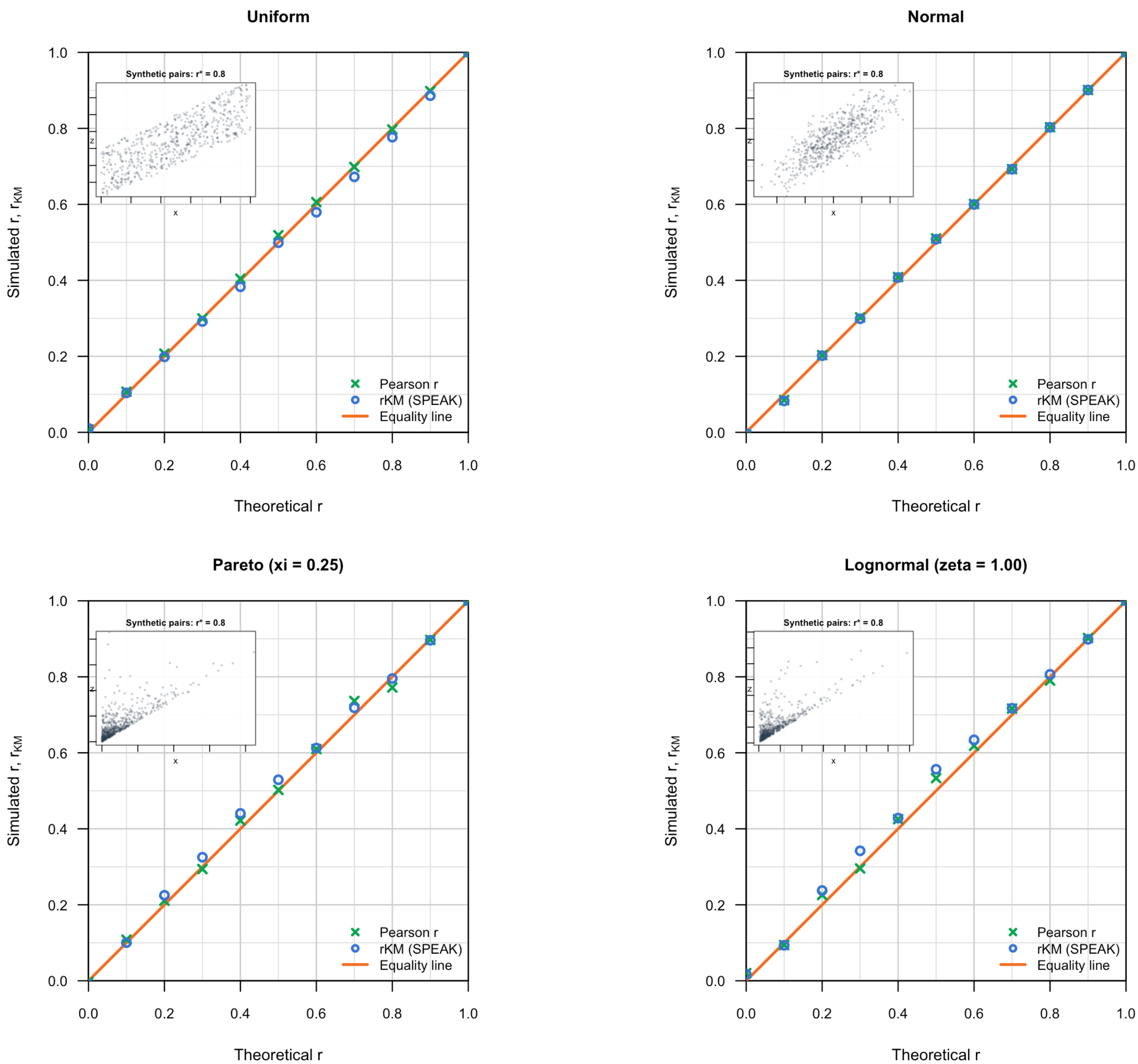

Figure R2.1. Synthetic experiments to compare standard correlation from Pearson (r) and

K-based correlation (r_{KM}).

Additionally, Figure R2.2 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.

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

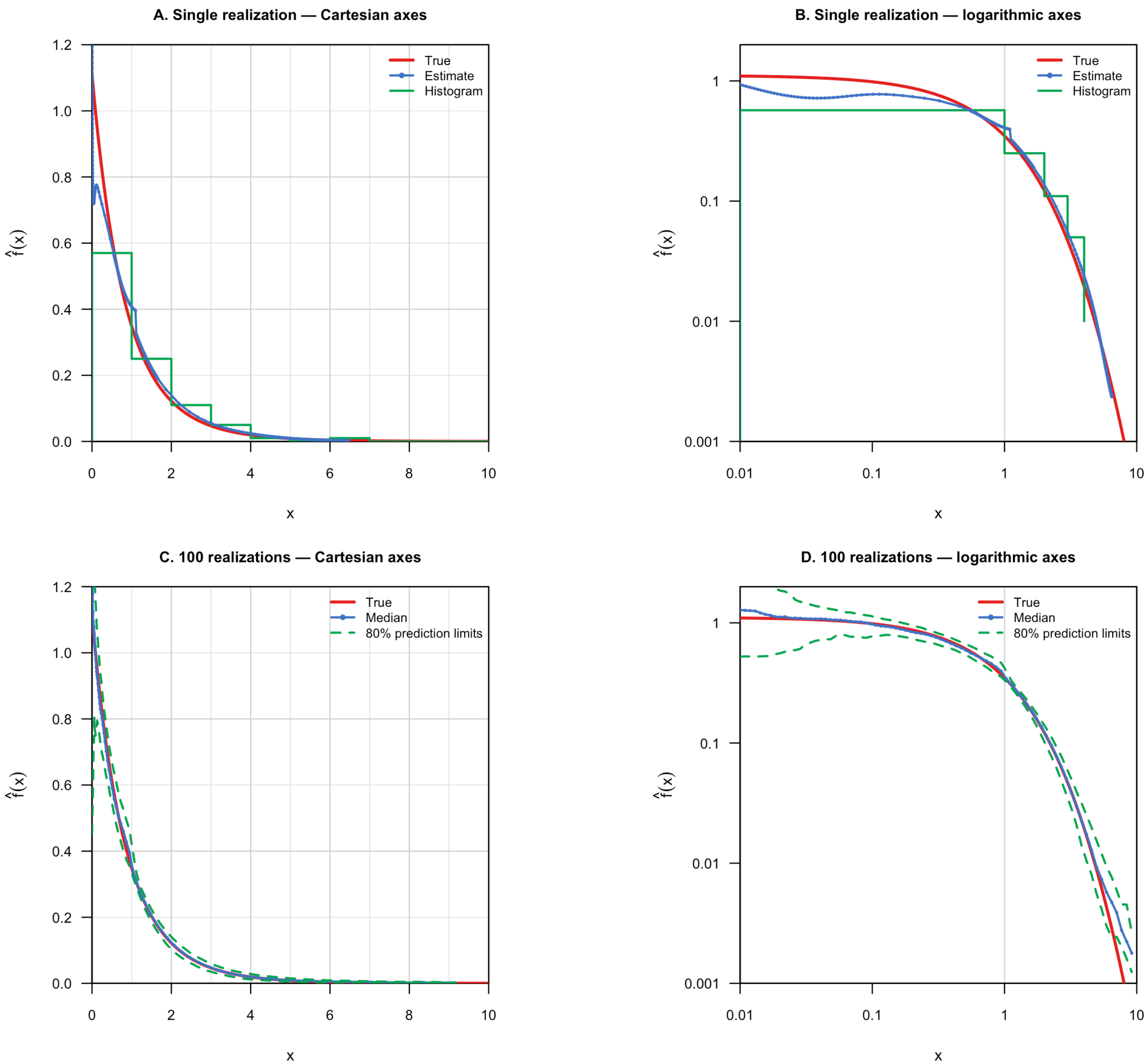

Figure R2.2. Synthetic experiments to illustrate KPDF estimation.

RC 4: Section 2.3: What does the variable Q represent in Equation (7)? This should be explicitly defined.

AR 4: Thank you for raising this point. It is a typo that will be fixed in the new manuscript version. Equation (7) reads now: $H(P_{ref}, P_{sim}) = \frac{1}{\sqrt{2}}\left(\sum_{x \in X}\left(\sqrt{P_{\mathrm{ref}}} - \sqrt{P_{\mathrm{sim}}}\right)^2\right)^{\frac{1}{2}}$

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

UNIVERSIDAD DEL BÍO-BÍO

RC 5: Section 3.1: The CAMELS-CL dataset provides only catchment-aggregated attributes. However, a spatially distributed hydrological model is applied. I therefore miss a description of the spatial datasets used to construct the distributed model. Only the CR2MET precipitation and temperature products are referenced. Please specify the data sources for elevation, soils, land use, and any other spatial information required by the hydrological model.

AR 5: Thank you for this observation. To run the TUW model, it only needs precipitation, air temperature, potential evapotranspiration, and topography (see Figure R2.3 below, which is the same as Figure 2 of the manuscript). Precipitation and air temperature use the CR2MET product, which is a national developed and distributed meteorological product. CR2MET has a horizontal resolution of 0.05° x 0.05° (regular grid). Potential evapotranspiration uses the Hargreaves-Samani equation and air temperature data (as a consequence, PET is also in a regular grid of 0.05° x 0.05°). Note that following recommendations from reviewers, the new used spatial resolution is 0.10°. That decision was made to ensure the native reference GLEAM data resolution and spatial patterns derived. See AR 5 (Reviewer 1) for new results at 0.10° spatial resolution, also incorporating WSPAEF as objective function for calibration purposes.

For topography, we used the 30 meters horizontal-resolution SRTM digital elevation model. No gridded soil, vegetation, or land-use dataset was used to regionalise TUW parameters.

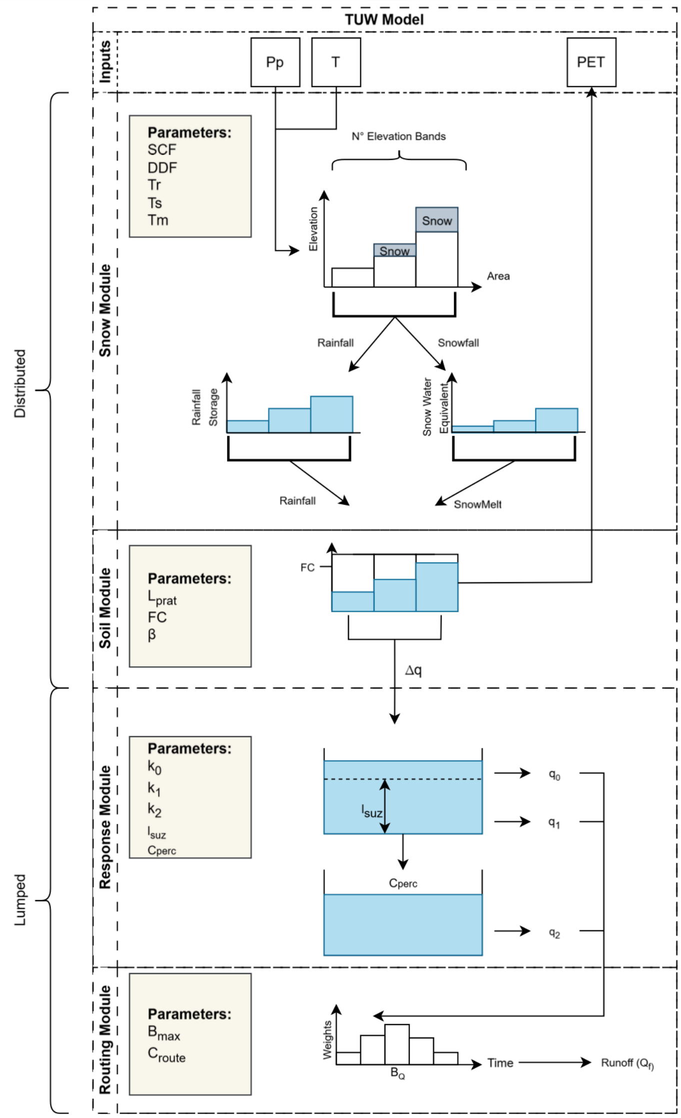

Figure R2.3. TUW rainfall-runoff model structure.

RC 6: Section 3.1: I also question whether bilinear interpolation is an appropriate method for resampling the GLEAM ET product. Bilinear interpolation introduces artificial spatial variability that is not present in the original data. A nearest-neighbour resampling approach would preserve the original observations. Alternatively, the simulated ET could be aggregated to the native GLEAM resolution before calculating the spatial performance metrics.

AR 6: 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 R2.1 shows the 43 catchments with 15 or more pixels at a spatial resolution of 0.10°:

Table R2.1. 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:

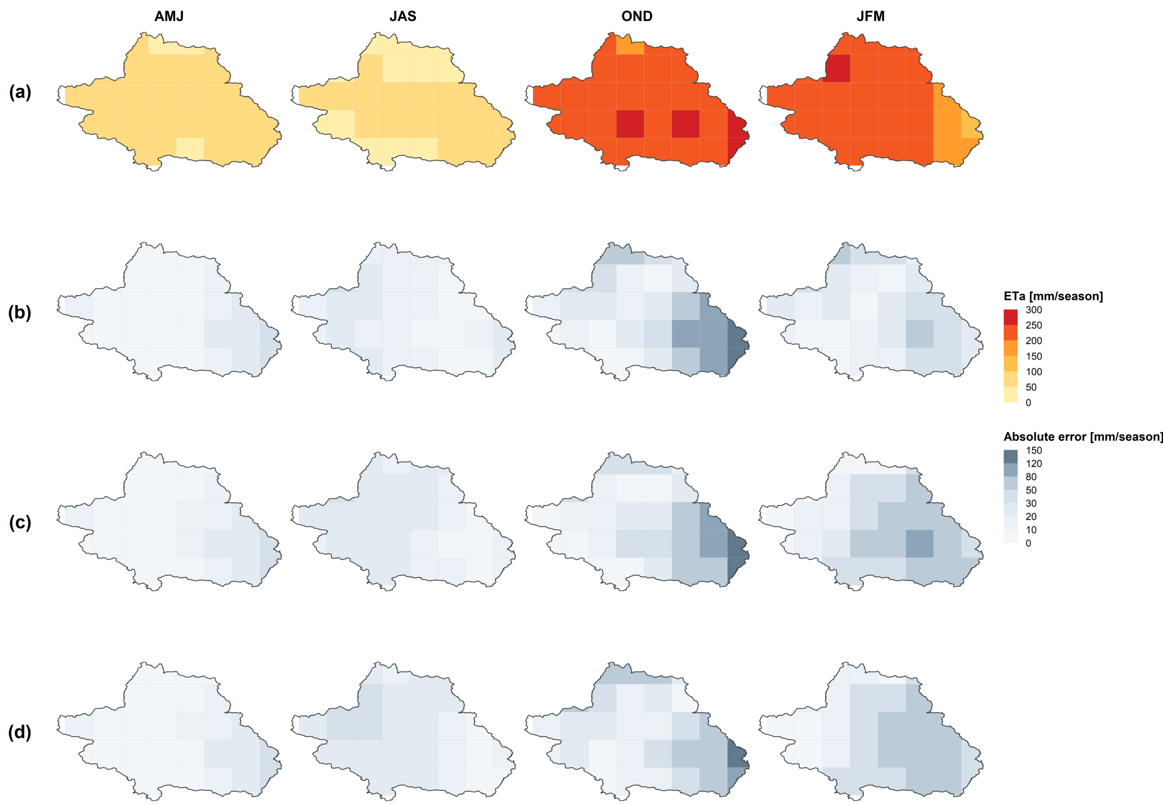

Figure R2.4 (Figure 4 in the main text): (a) Reference ETa from GLEAM for each three-month season (AMJ, JAS, OND, JFM), and spatial distribution of relative errors in simulated ETa with the TUW model with (b) the best SPEAK, (c) the best SPAEF, and (d) the best WSPAEF calibrations for catchment ID: 11302001 (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 R2.5 (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.

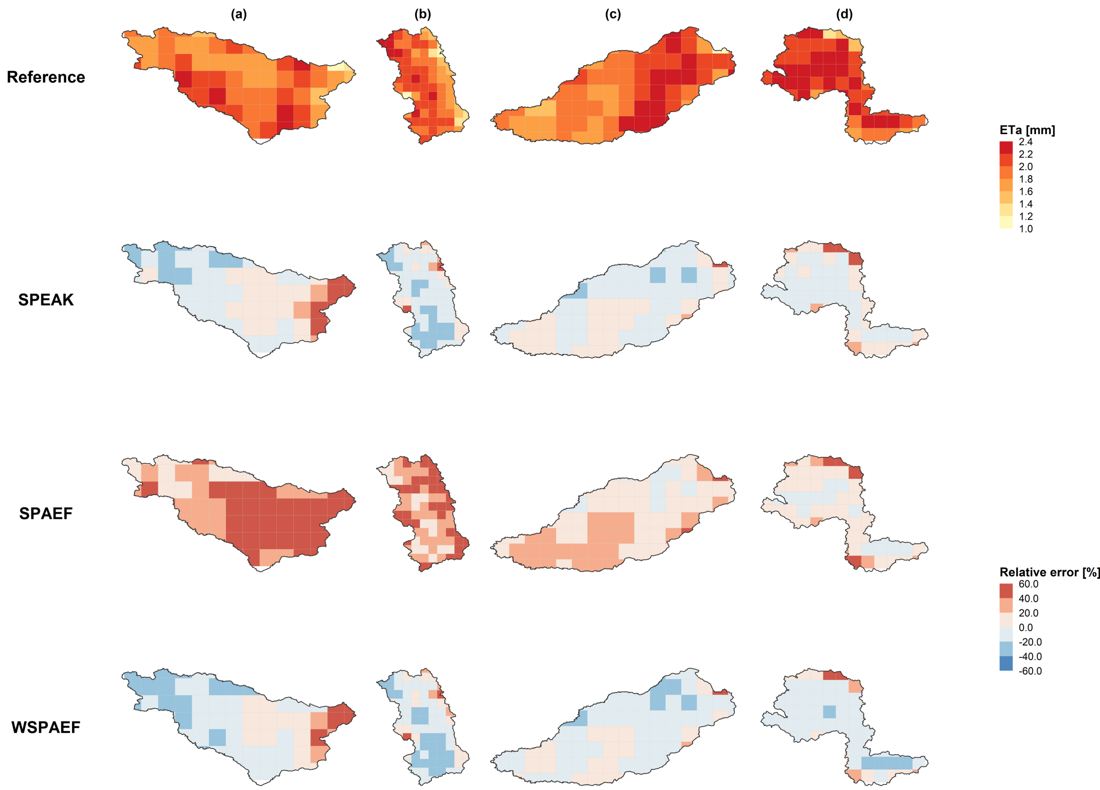

Figure R2.6 (Figure 6 in the main text): Comparison between GLEAM actual evapotranspiration fields (annual mean per pixel) and simulated ETa using parameters obtained by calibrating with SPEAK, SPAEF, and WSPAEF against GLEAM reference data. Each column corresponds to a specific catchment: (a) Río Itata 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 R2.2. Moran index and closer agreement with GLEAM data.

		MORAN Index				
Catchment ID	N° pixels at 0.10°	GLEAM	SPEAK	SPAEF	WSPAEF	Closer to GLEAM
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**	

d) Summary of results presented in AR 6.c:

Table R2.3. 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 7: Section 3.2: Please justify why modelling scenarios (b) and (c) are appropriate for evaluating the proposed metric

AR 7: The two scenarios were intended as controlled forcing-discretisation experiments while holding the TUW structure and catchment-level parameterisation constant. The elevation-band configuration preserves elevation-dependent snow processes; the gridded-forcing configuration preserves the 0.10° meteorological forcing patterns. We will add this rationale and avoid describing the second scenario as a fully distributed hydrological model. The main text will read now:

> … Scenarios (b) and (c) are used to compare the objective functions under two spatial representations of meteorological forcing while retaining the same conceptual model and one parameter vector per catchment. Scenario (b) represents forcings through 200 m elevation bands, whereas scenario (c) supplies forcings on the 0.10° CR2MET grid (aggregated to match the native GLEAM spatial resolution). These scenarios do not represent spatially distributed model parameters.…

RC 8: Section 3.2: Given that the objective is to assess spatial ET patterns, I am particularly surprised that spatially uniform model parameters are employed. ET patterns are expected to be strongly influenced by spatial variability in land use, vegetation, and soils. Consequently, I question whether the chosen model configuration is suitable for evaluating a spatial ET performance metric.

AR 8: We decided to use spatially uniform parameter sets to simplify the calibration experiments, obtaining spatial variability in model states and fluxes only due to spatial heterogeneities in meteorological fields. In the revised manuscript, we will correct the model description, rename the 'fully distributed' case, and remove process claims involving spatial soil, vegetation, or land-use controls. The following sentences will be added to the main text:

c) Fully distributed model configuration will be renamed as 'gridded-forcing configuration':

> … In the gridded-forcing configuration, precipitation, temperature, and PET were supplied on the 0.10° spatial resolution (i.e., native GLEAM resolution), while one calibrated TUW parameter vector was applied uniformly to all grid cells within each catchment. Thus, this configuration represents spatial variability induced by the meteorological forcings and model states, but not spatial heterogeneity in soil-, vegetation-, or land-use-dependent parameters. Testing SPEAK in a model with pedotransfer functions or parameter regionalisation remains an important next step.…

d) Discussion section with related information and new research paths:

> … Future research will attempt to test the potential of SPEAK using a spatial parameter configuration that considers pedotransfer functions or other spatial regularization strategies.…

RC 9: Section 3.3: Why is each catchment calibrated independently? One of the motivations for spatial performance metrics is to improve the simulation of large-scale spatial patterns. It would therefore seem more appropriate to calibrate a common parameter set capable of reproducing realistic ET gradients across multiple catchments.

AR 9: Each catchment was calibrated independently because TUW parameters are catchment-specific conceptual parameters. Therefore, the objective-function comparison is paired within each catchment. A single parameter set across Chile would combine the metric comparison with a national parameter-regionalisation problem and would not be a minimal or controlled extension. We will therefore retain independent calibration, clarify that the analysis concerns within-catchment spatial patterns, and identify common or hierarchical parameter sets as future work. We will add a paragraph in the discussion section:

> … Each catchment was calibrated independently so that SPEAK and the benchmark objective functions were compared with the same model flexibility and parameter bounds. This work evaluates within-catchment ETa patterns and paired performance across catchment. It does not estimate one national parameter vector or one continuous national ETa field. Multi-catchment regional or hierarchical calibration is outside the scope of this study.…

RC 10: Section 3.3: Please also clarify how the ET metrics were calculated. Were they computed from daily/monthyl ET maps, or long-term averages/climatologies? How were the ET maps aggregated temporally before calculating the metrics?

AR 10: We agree that the original description was insufficient. We will distinguish the daily space-time samples used by the calibration objective from the temporally aggregated fields used in the figures and out-of-sample evaluation. The revised text and every caption will state the dates, units, aggregation, mask, and any other relevant information. As an example, the following adding will be provisional considered:

> … During calibration, SPEAK and SPAEF were computed from paired daily ETa matrices after the warm-up period; the matrices were flattened over all valid day-pixel pairs. During evaluation (1 April 2002-31 March 2020), daily space-time metric values are distinguished from long-term mean and seasonal climatological ETa fields. Seasonal fields were formed for AMJ, JAS, OND, and JFM using identical day-pixel masks for reference and simulation. The temporal aggregation and units used in each figure are stated in its caption.…

RC 11: Section 3.3: Was the discharge KGE also included as an objective during the SPAEF and SPEAK calibrations?

AR 11: Discharge KGE was not included in the SPAEF or SPEAK calibrations. As a consequence, those calibrations were ETa-only.

RC 12: Section 3.3: Please specify whether the NRMSE and FSS metrics were calculated for both discharge and ET, or only for one of these variables.

AR 12: NRMSE and FSS were calculated only for ETa. Discharge was evaluated with KGE. We will state this at the first mention of the metrics and in the relevant figure captions. An ad-hoc explanation will be added to the main text:

> …NRMSE and FSS were computed only for ETa fields; discharge performance was evaluated separately with KGE.…

RC 13: Section 4: Figure 3: Panels (b) to (f) are missing units on the y-axis.

AR 13: Thank you. Units will be added to the y-axis in Figure 3.

RC 14: Section 4: Figure 4: Since SPAEF and SPEAK are designed to be insensitive to bias, I do not see the relevance of presenting absolute ET errors. It is commonly the discharge that is used for water balance closure at basin scale. Also, absolute values of ET products are associated with uncertainties but often contain useful spatial pattern information. This has motivated the definition of the bias-insensitive SPAEF metric.

AR 14: Thank you for raising this point again. Please, see AR1 where it is stated that SPEAK is not fully bias-insensitive (we separated magnitude and pattern diagnostics into components of MSE (see AR 1) and also computed the Moran Index).

We agree that absolute ETa error cannot be used as primary evidence of spatial-pattern skill. As a consequence, we separated magnitude and pattern diagnostics into components of MSE (see AR 1) and also computed the Moran Index. Some results complementing AR6 as follows:

	11141001		10304001		8334001	
	vs SPAEF	vs WSPAEF	vs SPAEF	vs WSPAEF	vs SPAEF	vs WSPAEF
Pearson	+0.210	+0.226	+0.109	+0.464	+0.078	+0.275
Spearman	+0.237	+0.275	+0.124	+0.566	+0.100	+0.302
RMSE_Z	+0.305	+0.324	+0.098	+0.369	+0.060	+0.202
MSE pattern component	+0.023	+0.025	+0.004	+0.022	+0.020	+0.025
Moran error	+0.116	+0.132	+0.093	+0.066	+0.059	-0.030
Standardized variogram error			+0.078	+0.089	+0.217	+0.451
NMI	+0.289	+0.215	+0.018	+0.033	+0.022	+0.011
Quantile-matched FSS	+0.031	+0.031	-0.012	+0.124	+0.096	+0.111

SPEAK better Competitor better

Figure R2.7: Illustration of three different size catchment comparisons among SPEAK, SPAEF, and WSPAEF. From left to right, catchments have 18, 30, and 150 pixels. Eight different assessment metrics considered (vertical axis). Green means SPEAK better, orange means competitor metric (SPAEF or WSPAEF) better. The shown number of catchments was decided for illustration purposes only.

UNIVERSIDAD DEL BÍO-BÍO

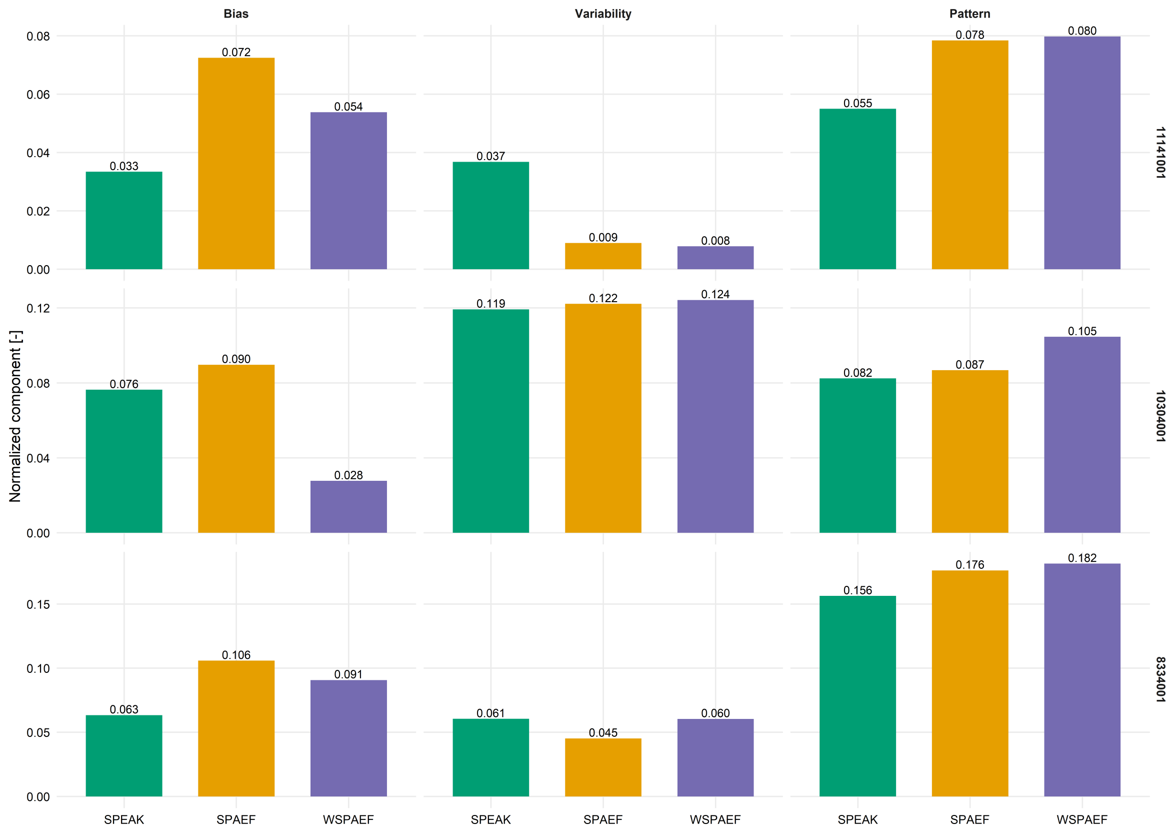

Figure R2.8: The MSE decomposition factors are shown for the same catchments in Figure R2.7. The shown number of catchments was decided for illustration purposes only.

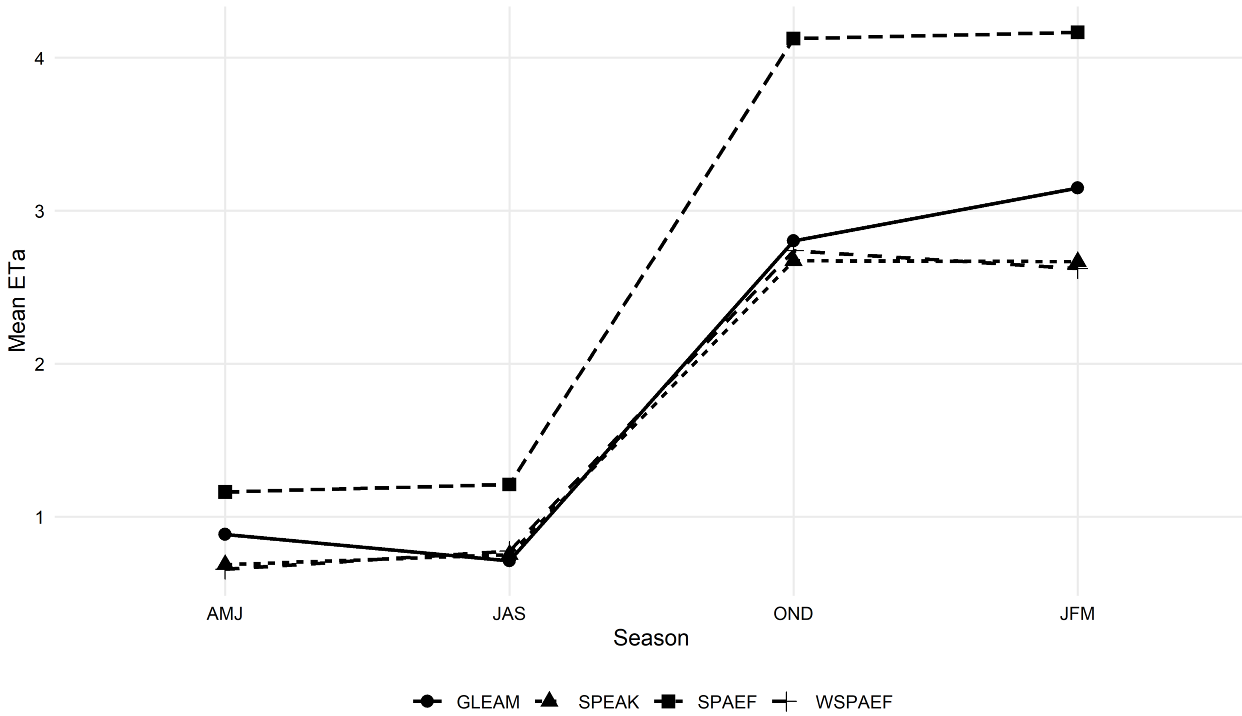

Figure R2.9: Mean seasonal ETa for catchment ID 8317001 (104 pixels), considering GLEAM (reference), SPEAK, SPAEF, and WSPAEF.

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

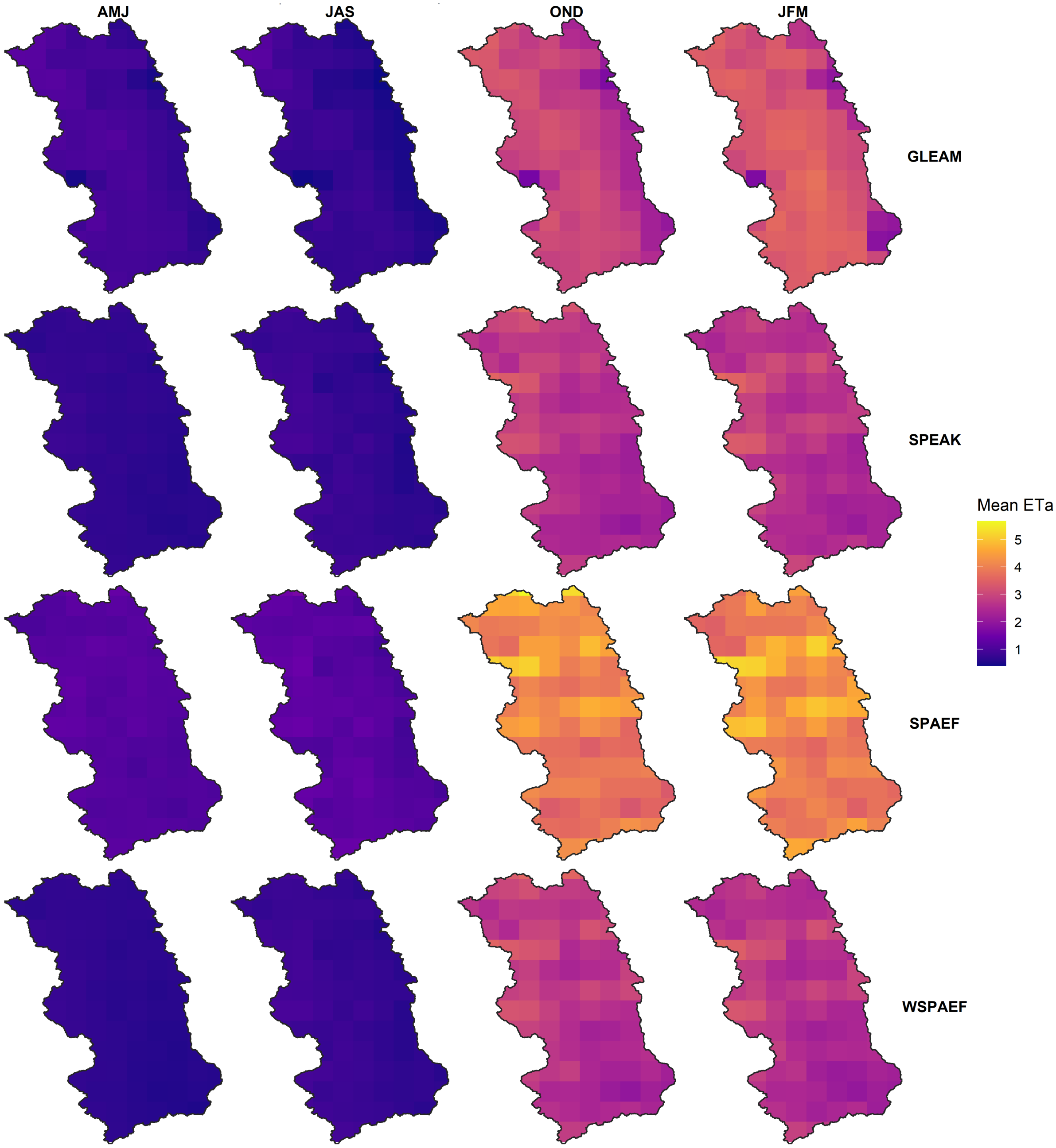

Figure R2.10: Mean (pixel by pixel) seasonal ETa for catchment ID 8317001 (104 pixels), considering GLEAM (reference), SPEAK, SPAEF, and WSPAEF.

UNIVERSIDAD DEL BÍO-BÍO

RC 15: Section 4: Figure 6: Similar to Figure 4, the discussion focuses on relative errors, although SPAEF and SPEAK are intentionally insensitive to bias. I therefore question the usefulness of these comparisons.

AR 15: See AR1, AR6, and AR14.

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