SUPPLEMENTARY MATERIAL

1. Data and Methodological Supplementary Material

Supplementary material is attached that expands on the data and the methodology presented in the main document.

1.1. Climate Change Models

Table S1 Preselected EURO-CORDEX climatic model ensembles

GCM	RCM	GCM	RCM
CCCma-CanESM2	SMHI-RCA	MIROC5	SMHI-RCA
CNRM-C5	CNRM-ALADIN	MOHC-HadGEM2	KNMI-RACMO
CNRM-C5	SMHI-RCA	MOHC-HadGEM2	SMHI-RCA
CSIRO-Mk3	SMHI-RCA	MPI-ESM	CLMcom-CCLM
ICHEC-EC-Earth	DMI-HIRHAM	MPI-ESM	MPI-REMO R1
ICHEC-EC-Earth	KNMI-RACMO	MPI-ESM	MPI-REMO R2
ICHEC-EC-Earth	SMHI-RCA	MPI-ESM	SMHI-RCA
IPSL-CM5A	INERIS-WRF	NCC-NorESM	SMHI-RCA
IPSL-CM5A	SMHI-RCA	NOAA-ESM	SMHI-RCA

1.2. Proposed PET estimation method

The Thornthwaite equation was applied together with a proposed modified version of the Marcos-Garcia (2019)[1] equation, which replaces the mean temperature parameter with a corrected effective temperature. The proposed modification includes a dual seasonal parametrisation for calculating the effective temperature parameter, distinguishing between the warmest and coldest seasons (Equation 1)

$$T' = a_s(T + A)^{1-b_s} \qquad \textbf{Eq. 1}$$

Where:

T': Effective temperature.

T: Mean temperature.

A: Monthly maximum temperature range.

a_s, b_s: Setting seasonal parameters

The estimation accuracy with the proposed modification was compared with four (lower data-intensive) estimation methods:

- *Thornthwaite*
- *Camargo et al. (1999)*[2]
- *Pereira & Pruitt (2004)*[3]
- *Marcos-Garcia (2019)*

The PET data-intensive Penman-Monteith estimations registered for the Teruel meteorological station in the basin were used as a reference for comparing the degree of seasonal adjustment.

[1] Marcos-García, P. (2019). Sistema de ayuda a la decisión para la adaptación y gestión de sistemas de recursos hídricos en un contexto de alta incertidumbre. Aplicación a la cuenca del Júcar [Doctoral Thesis]. Universitat Politècnica de València. https://doi.org/10.4995/Thesis/10251/125702

[2] Camargo, A.P., Marin, F.R., Sentelhas, P.C., Picini, A.G. (1999). Adjust of the Thornthwaite's method to estimate the potential evapotranspiration for arid and superhumid climates, based on daily temperature amplitude. *Rev. Bras. Agrometeorol.* 7 (2), 251–257

[3] Pereira, A.R., Pruitt, W.O. (2004). Adaptation of the Thornthwaite scheme for estimating daily reference evapotranspiration. *Agric. Water Manage.* 66, 251–257.

1.3. Hydrological model evaluation

Based on the multi-step criteria proposed by Krysanova (2018)[4], a Temez model suitability evaluation, mainly based on the Kling–Gupta Efficiency (KGE) indicator, was performed to assess:

Quality of observational data: The observed climate's spatial homogeneity and representativeness were assessed. The coverage of the entire calibration and validation period, as well as the spatial consistency of the observed hydrological data, were also evaluated.

Performance across climate patterns: The hydrological model's ability to reproduce the streamflow situation in exceptional weather conditions was assessed separately for the 10 warmest, 10 coldest, 10 wettest, and 10 driest years from 1961 to 2015 at the Manises gauge.

Multi-spatial performance: The hydrological model's ability to accurately reproduce streamflow in all regions was tested to prevent spatial bias in the evaluation.

Reproducibility of the indicator of interest: The skill of the hydrological model, forced with observed climate data, in reproducing the long-term inter-annual and seasonal streamflow shifts was evaluated.

1.4. Sub-selection methodology of Historical Performance-based methods

1.4.1. Best performing seasonal climate depiction (SCD)

This method selects the model that best reproduces the observed long-term monthly climatic mean patterns (precipitation and temperature) during the reference period (1961–1990). Performance is evaluated using the RMSE distance-based metric between simulated and observed climatological monthly means.

1.4.2. Best-performing seasonal hydrological depiction (SHD)

Similar to SCD, this method evaluates model performance based on the ability to reproduce observed long-term monthly mean streamflow patterns during the reference period. Selection is based on minimising the distance (RMSE) between simulated and observed seasonal streamflow behaviour.

1.4.3. Simple climate model weighing (CMW)

Models are ranked according to their climatic performance during the reference period, based on their ability to reproduce long-term monthly mean climatic variables. Models are included in a weighted ensemble in proportion to their performance ranking, so higher-performing models are replicated more frequently. From the resulting weighted ensemble, the median-performing model is selected.

1.5. Sub-selection methodology of Mixed methods

1.5.1. Climate GCM diversity (CGCM)

This method groups simulations by their driving Global Climate Model (GCM), thereby promoting structural diversity across large-scale climate drivers. From each GCM group, the member with the best climatic performance during the reference period (based on long-term monthly mean diagnostics) is selected.

1.5.2. Climate RCM diversity (CRCM)

Analogous to CGCM, this method groups simulations according to their Regional Climate Model (RCM) configuration. From each RCM group, the member with the highest climatic performance in reproducing long-term monthly mean variables during the reference period is selected.

[4] Krysanova, V., Donnelly, C., Gelfan, A., Gerten, D., Arheimer, B., Hattermann, F., & Kundzewicz, Z. W. (2018). How the performance of hydrological models relates to credibility of projections under climate change. *Hydrological Sciences Journal*, *63*(5), 696–720. https://doi.org/10.1080/02626667.2018.1446214

1.6. Sub-selection methodology of Reliability Ensembles sub-selections

This section describes the implementation of the three reliability-based sub-selection methods evaluated in this study: Reliability Ensemble Average (REA), Reliability Ensemble on Historical Performance (REHP), and Reliability Ensemble on Future Consensus (REFC). In all cases, model ranking is based on normalised reliability scores derived from performance or consensus diagnostics based on the Tebaldi & Knutty (2007) REA weighting method.

1.6.1. Normalisation of performance and consensus diagnostics

For a given model m and basin b, diagnostic metrics were transformed to a common 0–1 scale (with 1 indicating best performance) using inverse min–max normalisation:

$$S(x_{m,b}) = 1 - \frac{x_{m,b} - min_m(x_{m,b})}{max_m(x_{m,b}) - min_m(x_{m,b})}$$

where $x_{m,b}$ represents the diagnostic value (e.g. RMSE). This transformation ensures comparability across variables and maintains consistent interpretation (higher values indicate better reliability).

Future consensus was quantified as the absolute deviation of each model's projected change signal from the ensemble median:

$$C_{m,b} = \left|\Delta y_{m,b} - median_m(\Delta y_{m,b})\right|$$

where $\Delta y_{m,b}$ denotes the projected change in the indicator of interest. Consensus scores were then normalised using the same inverse scaling procedure so that models closer to the ensemble median receive higher scores.

1.6.2. Methods scoring

The reliability scores combine historical hydrologic and/or climatic performance, based on streamflow signal change, precipitation, and PET (as an integrating variable for average temperature and monthly maximum and minimum temperatures), with future consensus. For each basin, the model-specific reliability score is defined as follows.

$$W_{m,b}^{REA} = \frac{S(RMSE_{m,b}^{pr}) + S(RMSE_{m,b}^{PET}) + S(C_{m,b}^{\Delta stf})}{3}$$

Where $RMSE_{m,b}^{pr}$ and $RMSE_{m,b}^{PET}$ represent climatic historical performance during the reference period and $C_{m,b}^{\Delta stf}$ represents the *pseudo-future* consensus of projected streamflow change.

$$W_{m,b}^{REHP} = \frac{S(RMSE_{m,b}^{pr}) + S(RMSE_{m,b}^{PET}) + S(RMSE_{m,b}^{stf})}{3}$$

Where all components represent the ability to reproduce long-term monthly mean climatic and hydrological variables during the reference period. No future consensus information is incorporated.

$$W_{m,b}^{REFC} = \frac{S(C_{m,b}^{\Delta pr}) + S(C_{m,b}^{\Delta PET}) + S(C_{m,b}^{\Delta stf})}{3}$$

where each term represents the normalised distance of projected change signals from the ensemble median.

After obtaining the incremental values of all climate models through the scoring process of each Reliability sub-selection method, the models were ranked, and the Knee Point Detection and Jenks' natural breaks classification methods were applied to select the smallest best-performing sub-ensemble between the two selections. Knee Point Detection is a mathematical approach widely used in multi-objective optimisation to identify the point of maximum curvature, where marginal gains in a ranked metric decline notably. By contrast, Jenks' natural breaks classification, first formalised by Jenks (1967), is a clustering method that minimises within-class variance and maximises between-class variance.

1.7. Sub-selection methods summary

Table S2 Sub-selection methods summary

Method	**Category**	**Primary Basis of Selection**	**Uses Historical Performance?**	**Uses *Future* Information?**	**Selection Logic**	**Key Idea**
SCD	Performance-based	Climatic variables (pr, tas, tasmax, tasmin)	Yes	No	Lowest RMSE of long-term monthly mean climate	Select the model best reproducing the observed seasonal climate
SHD	Performance-based	Streamflow	Yes	No	Lowest RMSE of long-term monthly mean streamflow	Select the model best reproducing the observed hydrological behaviour
CMW	Performance-based (weighted ranking)	Climatic variables	Yes	No	Performance-based ranking and weighted ensemble selection	Weight models by climatic skill and select the median model
CGCM	Diversity-based (structural)	GCM grouping, Climatic variables	Yes	No	Select the best-performing member within each GCM family	Preserve large-scale structural diversity
CRCM	Diversity-based (structural)	RCM grouping, Climatic variables	Yes	No	Select the best-performing member within each RCM family	Preserve regional structural diversity
REA	Reliability-based	Climate performance + future streamflow consensus	Yes	Yes	Average normalised reliability score	Balance historical skill and future agreement
REHP	Reliability-based	Climate + streamflow performance	Yes	No	Average normalised historical reliability score	Select models with the highest historical reliability
REFC	Reliability-based	Future consensus (climate + streamflow change)	No	Yes	Average normalised consensus score	Select models closest to the ensemble *pseudo-future* consensus
KCRC	Statistical clustering (historical)	Climate patterns (reference period)	Yes	No	K-means clustering on historical climatology	Promote diversity based on historical climate similarity
HCRC	Statistical clustering (historical)	Climate patterns (reference period)	Yes	No	Hierarchical clustering on historical climatology	Select representatives across the historical similarity space
KCFC	Statistical clustering (future)	Projected change patterns	No	Yes	K-means clustering on change signals	Promote diversity in projected future responses
HCFC	Statistical clustering (future)	Projected change patterns	No	Yes	Hierarchical clustering on change signals	Select representatives across the projection similarity space

1.8. Methodology of K-means clustering on reference climatology sub-selection (KCRC)

The K-means clustering method is an unsupervised learning technique that groups a data set into k clusters based on the similarity between observations. The algorithm (**Fig. S1**) assigns each point to the randomly initialised nearest centroid, recalculates the centroids as the mean of the points in each group, and repeats this process until the centroids stabilise. It is an efficient and widely used method, although it requires a predefined number of clusters and can be sensitive to data scale and initialisation. Since the technique requires a random initialisation of the cluster centroids, 2000 runs were performed to mitigate the sensitivity to the initialisation point of the centroids. Iterative clustering was performed with k number of clusters where $k \in [1, N] \cap \mathbb{Z}$ with N being the number of climate models. Then, from the set of N sub-selections, the one that best represented the seasonal behaviour of the reference climate in each sub-basin was selected.

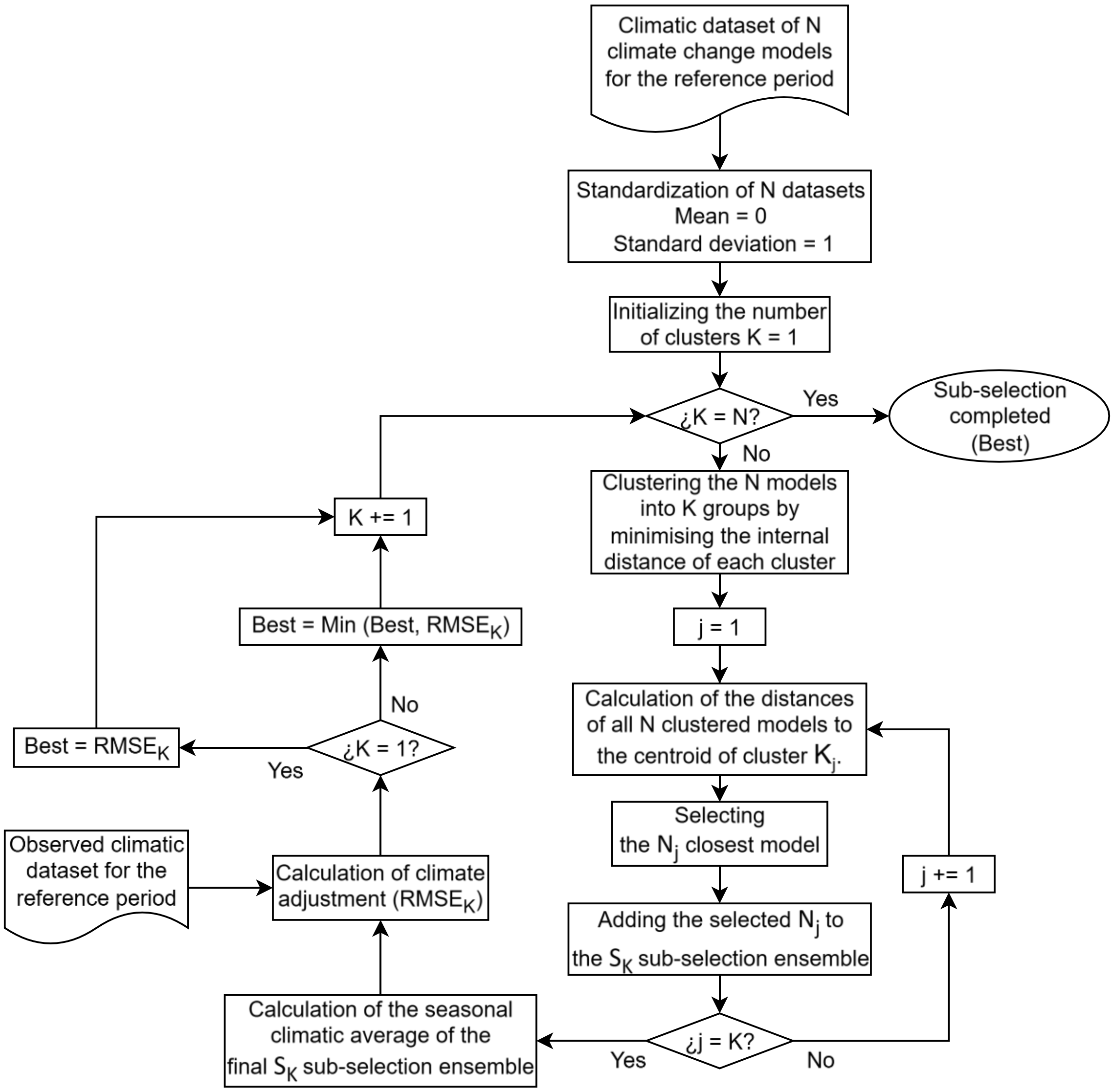


Fig. S1 KCRC Methodology

1.9. Methodology of agglomerative hierarchical clustering on reference climatology sub-selection (HCRC)

Hierarchical clustering is a grouping technique that constructs a hierarchy of similarity between elements, represented graphically by a dendrogram (**Fig. S2**). Each model starts as a member of its cluster. Based on Ward's method, the most similar clusters are merged to form a new cluster. The merging process continues with the number of clusters decreasing until all models are part of the same cluster. From the dendrogram, clusters can be identified by truncating the tree at a given level and dividing the set into a desired number of mutually exclusive clusters. This technique is advantageous when preserving information about the hierarchical relationships between the analysed elements is desired.

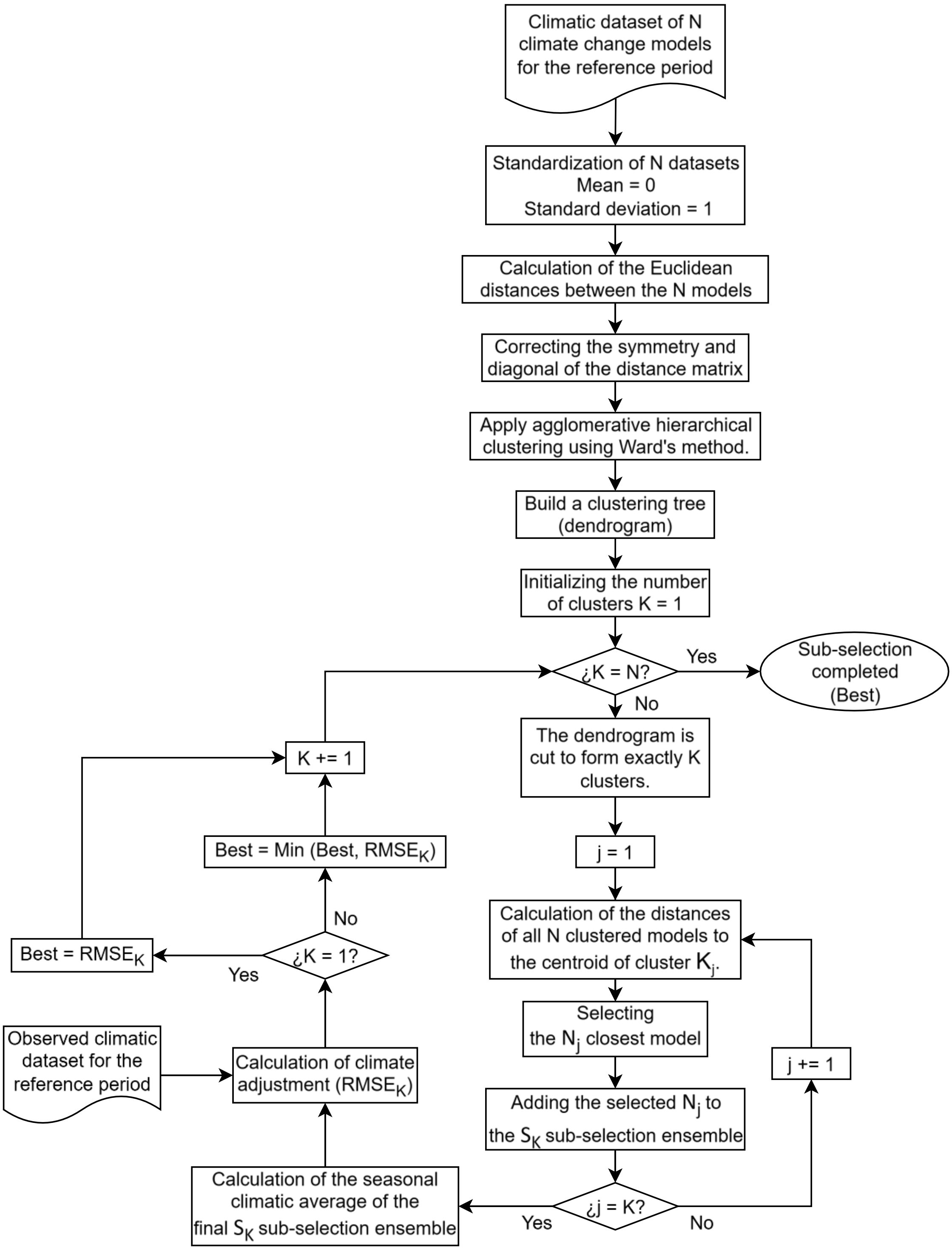


Fig. S2 HCRC Methodology

1.10. Methodology of K-means clustering on future change sub-selection (KCFC)

Similar to KCRC (1.8), but after the iterative clustering, the sub-ensemble selected is the one with the smallest size and the highest percentage of future streamflow change coverage regarding the no-selection ensemble spread (**Fig. S3**).

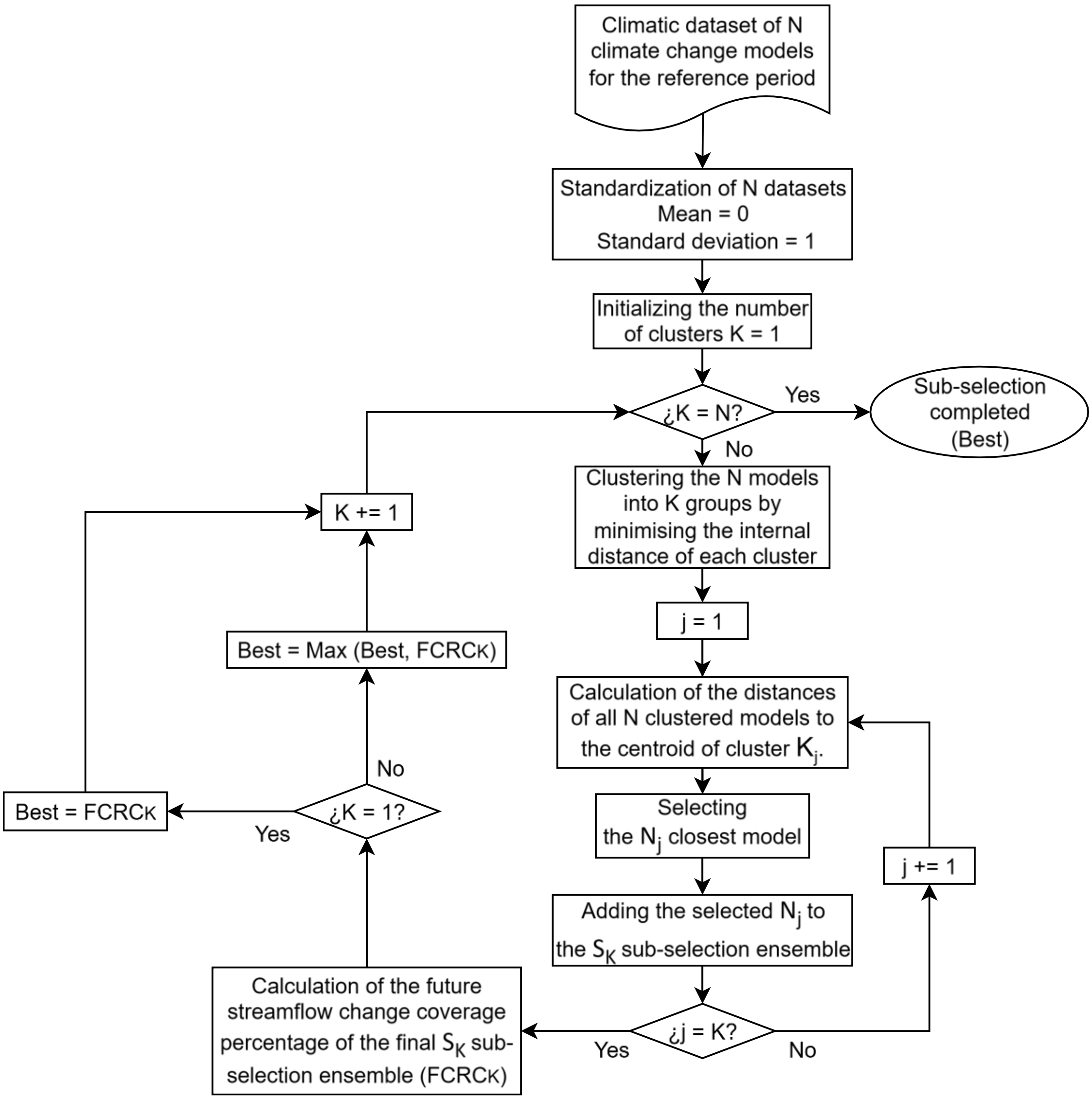


Fig. S3 KCFC Methodology

1.11. Methodology of agglomerative hierarchical clustering on future change sub-selection (HCFC)

Similar to HCRC (1.9), but after the iterative clustering, the sub-ensemble selected is the one with the smallest size and the highest percentage of future streamflow change coverage regarding the no-selection ensemble spread (**Fig. S4**).

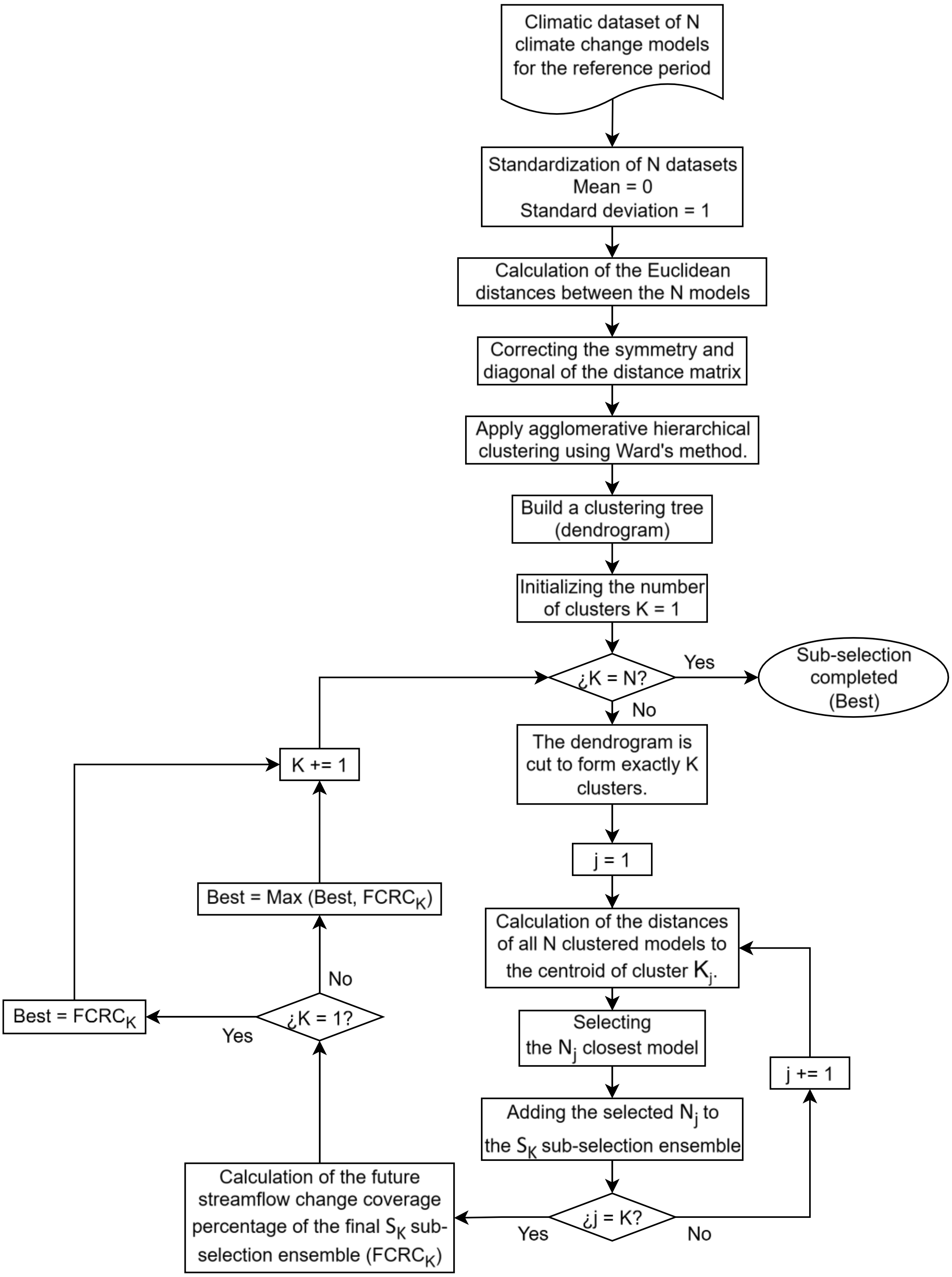


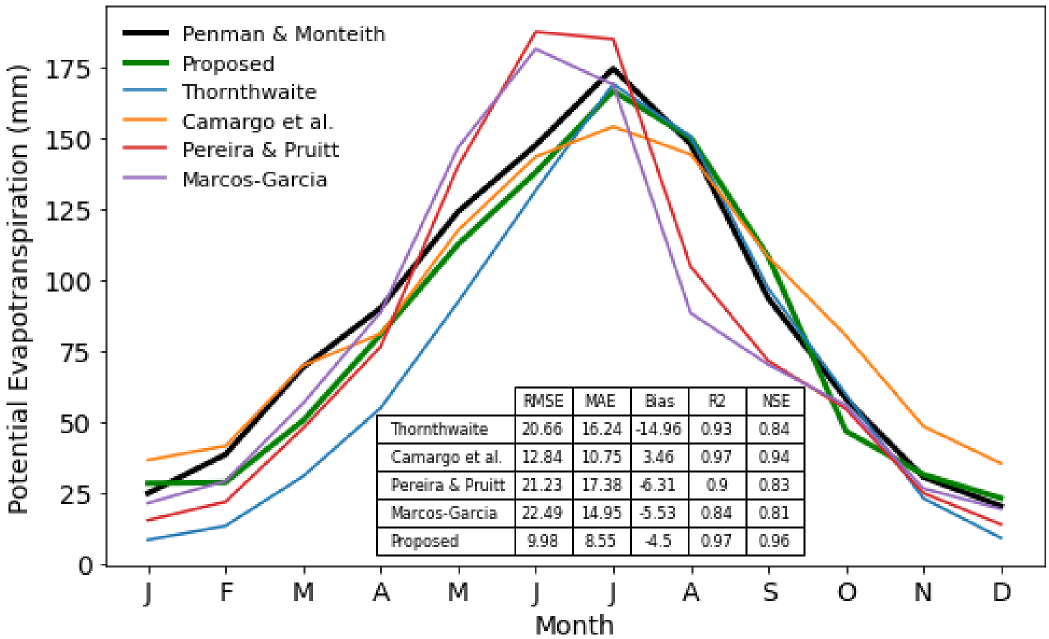
Fig. S4 HCFC Methodology

2. Results' Supplementary Material

Supplementary material is attached that expands on the results presented in the main document.

2.1. PET calculation method assessment

Figure S5 presents a seasonal comparison of PET estimation for the assessed methods. Based on multiple evaluation metrics, the proposed method achieves the best overall performance across all metrics, demonstrating both accuracy and stability. *Camargo et al.* also perform well, showing a small positive bias and strong correlation, although with slightly higher errors. By contrast, the original Thornthwaite formulation systematically underestimates potential evapotranspiration, as reflected in its large negative bias and higher RMSE, despite maintaining a reasonable correlation with observations. *Pereira & Pruitt* and *Marcos-Garcia* achieve moderate agreement but tend to underestimate values, showing weaker NSE scores. Overall, the proposal and *Camargo et al.* stand out as the most reliable methods.



	RMSE	MAE	Bias	R2	NSE
Thornthwaite	20.66	16.24	-14.96	0.93	0.84
Camargo et al.	12.84	10.75	3.46	0.97	0.94
Pereira & Pruitt	21.23	17.38	-6.31	0.9	0.83
Marcos-Garcia	22.49	14.95	-5.53	0.84	0.81
Proposed	9.98	8.55	-4.5	0.97	0.96

Fig. S5 Seasonal PET at the Teruel meteorological station

2.2. Hydrological model performance evaluation

Quality of observational data: The monthly climate dataset of observed precipitation and temperature from the SPAIN02 initiative has a high spatial resolution of 0.2° over the Iberian Peninsula, allowing it to adequately capture regional and sub-regional-scale climate patterns. The representation of local microclimates and their temporal accuracy for the period 1950-2015 is highly beneficial in the Turia River basin, thanks to kriging methodologies that utilise information from more than 12 meteorological stations. Regarding the quality of the hydrological information, the hydrological model was compared with the PATRICAL model dataset, which incorporates data from the Spanish gauge station network and calculates historical series under natural regime conditions from 1960 to the present. In conclusion, it is considered that the input information to the hydrological model is of high quality and adequate for the proposed purposes.

Table S3 Performance statistics (KGE, r = correlation, β = bias ratio, γ = variability ratio; all dimensionless; ideal value at unity) for the full time series of the 5 gauges and for the 10 warmest, 10 coldest, 10 wettest, and 10 driest years of the 55-year period at Manises Gauge near Valencia city

Gauge		KGE	r	β	γ
Arquillo de San Blas	55	0.77	0.82	1.00	0.85
Benageber	55	0.78	0.81	1.00	0.88
Loriguilla	55	0.85	0.85	1.00	1.00
Pueblos Castillo	55	0.82	0.83	0.98	1.05
Manises	55	0.90	0.91	0.97	1.00
Manises	10 warmest	0.84	0.95	0.98	0.85
Manises	10 coldest	0.86	0.86	1.02	1.02
Manises	10 wettest	0.85	0.93	1.05	0.87
Manises	10 driest	0.83	0.93	1.07	1.07

Exceptional weather conditions: The model demonstrated robust skill under climatic extremes, with KGE values ranging from 0.84 to 0.86 during the warmest, coldest, and wettest years, and 0.83 for the driest years. Correlation (r = 0.86–0.95) remained high, while bias (β ≈ 1.0) indicated accurate mean flow reproduction. Variability errors (γ = 0.85–1.07) exhibited slight underestimation during warm/wet periods and modest overestimation during dry years, but overall performance remained stable (**Table S3**).

Multi-spatial performance: Across sub-basins, KGE values ranged from 0.77 at Arquillo to 0.90 at Manises, with correlation consistently above 0.81. Bias ratios remained close to unity (β = 0.97–1.00), and variability (γ = 0.85–1.05) confirmed reliable reproduction of streamflow amplitude. Performance was highest in Mediterranean sub-basins (Loriguilla, Manises) and slightly weaker in continental headwaters, though without evidence of systematic spatial bias, indicating adequate model transferability across heterogeneous hydrological regimes.

Table S4 Performance statistics for the seasonal streamflow of the analysis periods and the seasonal streamflow change between them

Variable	KGE	r	β	γ
Streamflow (1961-1990)	0.83	0.86	1.01	1.09
Streamflow (1991-2015)	0.68	0.85	0.96	0.73
Streamflow change	0.76	0.84	1.18	1.03

Hydrological indicator reproducibility: **Table S4** illustrates how the model, forced with observed climate data, reproduced the long-term interannual streamflow regime with satisfactory skill during 1961–1990 (KGE = 0.83), supported by a high correlation (r = 0.86) and minimal bias (β = 1.01). Performance decreased in 1991–2015 (KGE = 0.68), primarily due to underestimated variability (γ = 0.73), yet correlation remained high (r = 0.85). Additionally, the model captured the directional shift in seasonal streamflow between periods (KGE = 0.76), reproducing the magnitude of reductions in summer and autumn, although with a tendency to overestimate the reduction change in November. These results confirm that the model adequately represents both historical conditions and observed hydroclimatic transitions, supporting its suitability for subsequent scenario-based analyses.

2.3. Frequency of model selection across basins and methods

The **Table S5** summarises the frequency with which each GCM–RCM combination was selected across all sub-selection methods and basins. This analysis provides an additional robustness check by identifying whether particular simulations influence the decision space. The results indicate that selection frequencies are distributed across models, with CNRM-C5_CNRM-LADIN and IPSL-CM5A_INERIS-WRF among the least frequently selected members. This suggests that its inclusion does not systematically drive sub-selection outcomes and that the framework naturally penalises simulations with comparatively lower relative performance or agreement.

Table S5 Frequency of model selection across basins and methods

Model Combination (GCM_RCM)	Basin						No. of times selected	
	AR	AB	BL	LP	PM	TUR	Total	Mean
MOHC-HadGEM2_KNMI-RACMO	10	6	3	6	6	6	37	6.2
CCCma-CanESM2_SMHI-RCA	3	9	3	6	4	9	34	5.7
MPI-ESM_MPI-REMO-R2	5	7	3	3	6	6	30	5.0
MPI-ESM_SMHI-RCA	2	3	8	7	8	2	30	5.0
ICHEC-EC-Earth_KNMI-RACMO	5	6	3	4	7	3	28	4.7
MIROC5_SMHI-RCA	3	4	4	5	7	3	26	4.3
CSIRO-Mk3_SMHI-RCA	4	3	6	4	5	2	24	4.0
ICHEC-EC-Earth_SMHI-RCA	4	3	3	4	5	3	22	3.7
NOAA-ESM_SMHI-RCA	3	2	4	6	4	3	22	3.7
IPSL-CM5A_SMHI-RCA	5	4	2	4	3	3	21	3.5
MPI-ESM_MPI-REMO-R1	3	3	5	1	6	2	20	3.3
NCC-NorESM_SMHI-RCA	2	3	5	3	3	3	19	3.2
ICHEC-EC-Earth_DMI-HIRHAM	3	4	2	0	5	3	17	2.8
MPI-ESM_CLMcom-CCLM	4	2	2	3	3	2	16	2.7
MOHC-HadGEM2_SMHI-RCA	0	1	4	6	4	1	16	2.7
CNRM-C5_SMHI-RCA	3	1	4	1	3	3	15	2.5
IPSL-CM5A_INERIS-WRF	2	2	3	1	3	1	12	2.0
CNRM-C5_CNRM-ALADIN	2	3	2	3	1	0	11	1.8