Dimensionality Reduction of the CORDEX-CA GCM-RCM Multimodel-ensemble on Precipitation using Principal Component Analysis (PCA) and Hierarchical Clustering (HC)

Maikel Mendez[1*], *Luis-Alexander* Calvo-Valverde[2], and *José-Andrés* Araya-Obando[1]

[1]Escuela de Ingeniería en Construcción, Instituto Tecnológico de Costa Rica, 159-7050, Cartago, Costa Rica
[2]Escuela de Ingeniería en Computación, Instituto Tecnológico de Costa Rica, 159-7050, Cartago, Costa Rica

Abstract. Principal Component Analysis (PCA) and Hierarchical Clustering (HC) were applied to reduce the dimensionality of a 19-member multimodel-ensemble combining different General Circulation Models (GCMs) and Regional Climate Models (RCMs) as part of the Coordinated Regional Climate Downscaling Experiment (CORDEX) for the Central America domain (CA). A subset of 12 Expert Team on Climate Change Detection and Indices (ETCCDI) was selected to evaluate the performance of each ensemble-member on precipitation against daily observational data from the Juan Santamaría International Airport (SJO), located in Alajuela, Costa Rica for the baseline period 1971-2000. The ETCCDI indices are designed to measure and quantify climate variability and associated trends. Results from the PCA analysis indicate that over 95% of the variance can be explained by the first three principal components (PC-1 through PC-3), showing high correlations, strong contributions and fair representation of most ETCCDI indices. HC clustering on the other hand, groups ensemble-members into 4 closely related clusters of common attributes (cluster-1 through cluster-4), with models ranging from dry to wet patterns. Afterwards, ensemble-members were sampled from each cluster to generate a sub-ensemble of representative simulations, reducing the original ensemble from 19 to 5 members, while still retaining its fundamental characteristics. Later, two multi-model ensemble-means (MEMs), one using the entire ensemble and the other using the 5-member subset were generated and their performance evaluated by means of five objective functions (nRMSE, MBE, MDA, PBIAS and MAE) against the observational dataset for the reference period. Nevertheless, no significant difference was found between both MEMs, implying that the applied techniques are effective in reducing dimensionality, preventing double-counting of highly dependent simulations, and consequently reducing the associated computational costs. Ultimately however, both MEMs noticeably overestimate seasonal precipitation during the reference period, suggesting the need for applying bias correction (BC) techniques prior to their use in impact assessment studies at local levels.

* Corresponding author : mamendez@itcr.ac.cr

1 Introduction

General Circulation Models (GCMs), refined with Regional Climate Models (RCMs) are the most advanced instruments available to study the climate system response to increases in radiative forcing and to identify the mechanisms driving such responses [1]. GCMs outputs however, are stacked with uncertainties relative to long-term historical observations, which typically arise due to systematic and random biases related to parameterization, boundary conditions and the physics-structure used in driving each model [2]. This issue is exacerbated since RCMs are usually nested inside GCMs, adding an additional layer of uncertainty [3]. Boundary conditions for multiple RCMs are often provided by only a few GCMs, while others are used only once. In contrast, comparable GCM boundary conditions generate significant interdependencies among RCM outputs, leading to even more ambiguity [4]. These inherent GCM-RCM uncertainties are frequently examined through the use of multi-model ensembles (MMEs). This challenges the assumption that multi-model climatological input will serve as the foundation for climate change impact assessments at local scale [5]. Additional complexities occur as MMEs cannot be considered unbiased, which is commonly the result of sample and model interdependence issues that affect both the mean and the variance of climate change signals [6]. Therefore, a thoughtful selection of climate simulations is required as input for climate change impact studies to avoid double-counting of similar model-members, reduce computational costs and mitigate biases in the ensemble statistics [7]. The objective of MMEs design should be to maximize model diversity to accurately capture model uncertainty and guarantee optimal model performance [8]. The same principles apply when choosing a sub-ensemble from a larger ensemble, where the conservation of statistical properties of the climate change signals should be reasonably represented by a balanced and unbiased subset of models [9]. Sampling model-members from a larger MME normally considers the selection of various climatic-relevant parameters (e.g. precipitation, temperature, relative humidity, etc.) based on complete historical observational datasets during a certain baseline period [10]. A more effective approach nonetheless, is the utilization of predefined indices, which adequately describe the main statistical features of climate variables in a compact form and are at the same time, reasonable in terms of detecting climate change signals [11]. Consequently, this study presents an approach to reduce the dimensionality of a multi-model ensemble (MME) built upon precipitation climate simulations from the Coordinated Regional Climate Downscaling Experiment (CORDEX) for the Central America domain (CA) based on the utilization of the Expert Team on Climate Change Detection and Indices (ETCCDI) for a case study in Costa Rica aimed to: (a) reduce sample biases, (b) appropriately account for uncertainties, and (c) reduce computational costs. This work is divided into multiple sections. First, ETCCDI indices are applied to both observations and CORDEX-CA simulations during the baseline period 1971-2000. Second, Principal Component Analysis (PCA) is used to determine the primary components of the MME. Third, Hierarchical Clustering (HC) is applied to identify closely related clusters of common attributes. Fourth, sampling criteria is applied to generate a subset of representative simulations. Fifth, two multi-model ensemble-means (MEMs), one using the entire ensemble and the other using the subset are generated and their performance evaluated by several metrics against the observational dataset for the reference period.

2 Methodology

2.1 Study area and observational datasets

Daily precipitation data from the Juan Santamaría International Airport (SJO) weather station in Alajuela, Costa Rica (Lat: 9.997, Long: -84.201), provided by Instituto Meteorológico of Costa Rica (https://www.imn.ac.cr/web/imn/inicio/) for the reference period 1971-2000 were used in this work. The city of Alajuela (Figure 1. (a)), lies approximately 60 km inland from the Pacific Ocean and 100 km from the Caribbean Sea at roughly 920 masl.

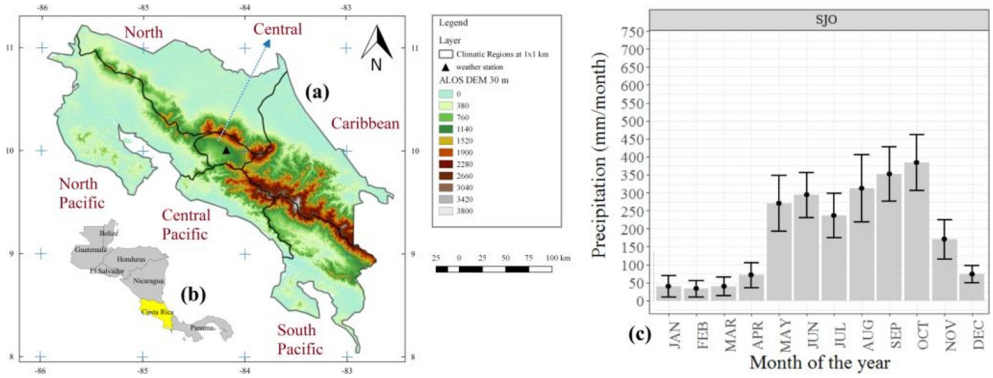

Fig. 1. (a) Location of SJO weather station and climatic regions in Costa Rica. (b) Position of Costa Rica in Central America and (c) seasonal distribution of precipitation during the reference period 1971-2000.

The city, with an area of approximately km2 located in the Central-Valley region of Costa Rica is surrounded by north-west and south-east cordilleras of high complexity and elevation, which promote oceanic climatological influences from both oceans [12]. Precipitation peaks in summer (JJA) and autumn (SON) (Figure 1. (b)) mostly derived from convective and frontal events resulting from a combination of climatic drivers including: (a) seasonal migrations from the Intertropical Convergence Zone (ITCZ), (b) sea breeze effects, (c) monsoon circulations, (d) strong easterly trade winds, (e) cold mid-latitude air masses in winter, and (f) disruptive influences of tropical cyclones in the Atlantic and Pacific Oceans [13]. The conjunction of all these factors considerably increases the vulnerability to flood risk for the city of Alajuela, exacerbated by rapid urbanization and outdated urban drainage systems. Consequently, the SJO weather station was selected as case study for this work on the basis of several factors including: (a) length and quality of historical records, (b) limited missing values for the reference period (< 1%), (c) geographical location and proximity to the city of Alajuela, and (d) relevance for the country's air transport sector.

2.2 Simulated datasets and climatic indices

Historical simulations were provided by the Coordinated Regional Climate Downscaling Experiment (CORDEX) for the Central America domain (CA) (https://esg-dn1.nsc.liu.se/projects/cordex/) covering a 19-member multi-model ensembles (MMEs) combining various GCMs and RCMs with spatial resolutions of 0.22° x 0.22° (~25 km) and 0.44° x 0.44° (~50 km) (Table 1). Daily simulated precipitation data from the nearest-to-station neighbouring grid-cell centre were extracted from each ensemble-member. Instead

of using the entire precipitation datasets for the reference period, a subset of 12 Expert Team on Climate Change Detection and Indices (ETCCDI) were selected to characterize the precipitation dynamics at the SJO weather station and to evaluate the performance of each ensemble-member against observational data (Table 1). The use of the well-established ETCCDI indices, which are reasonable in terms of detecting climate change signals, adequately describes the main statistical features of precipitation-patterns in a compact and efficient manner [14].

2.3 PCA Principal component analysis

Given the complexity of the CORDEX-CA array, Principal Component Analysis (PCA) was selected to: (a) reduce the dimensionality of the ensemble, (b) remove collinearities and redundancies among members, and (c) reduce random noise. PCA is a widely used approach to find low dimensional and interpretable representations of data that are inherently embedded in high-dimensional spaces [15]. PCA aims to explain as much of the original variance as possible by generating a new set of comprehensive and uncorrelated variables that are a linear combination of the initial variables by means of orthogonal transformation [16]. This new set of variables, are known as Principal Components (PCs), the Loadings or Coefficients of the linear combinations, indicate the relative importance of the original variables within the PCs [17]. In terms of procedure, PCA involves: (a) computing the covariance matrix of sample data, (b) calculating the eigenvalues and eigenvectors of this covariance matrix and (c) calculating the cumulative contribution rate [18]. In this study, PCA was applied to ETCCDI indices for each ensemble-member rather than using raw simulations. All indices were standardized to eliminate the influence of single-sample data before PCA.

2.4 Hierarchical clustering

Hierarchical clustering (HC) seeks to identify groups of simulations based on their behaviour in relation to common patterns previously found by PCA [19]. Specifically, Ward's minimum variance clustering (CA) was used in this study to discover groups within the CORDEX-CA ensemble based on the precipitation ETCCDI indices under characteristics of the reference period. Ward's method approaches cluster analysis as a variance problem by calculating Euclidean distances to assess the degree of dissimilarity between the clusters [20]. CA tends to find compact and spherical clusters of data, resulting in a tree-like similarity structure, which is meaningful given the fact that some clusters are more closely related to others. The optimal number of clusters was defined based on the broken-stick method, which compares the theoretical variances of a randomly generated datasets with the variances of individual PCs in the used dataset, cutting the tree where this increase of variance does not change considerably. This is carried out to lower noise and strengthen the ensuing cluster analysis.

2.5 Model similarity sampling

In this study, one simulation from each cluster of related models derived from the hierarchical clustering (HC) analysis was chosen using the quota-sampling method, where members of the ensemble are chosen from each group based on salient features or information pertinent to the subject under investigation. This is done by picking simulations from the principal component space plot of the PCs, starting with an average simulation where all PCs are closer to 0 and then, subsequently select one simulation from each cluster exhibiting unique and extreme characteristics [7]. As a result, the quota-sampling method

favors selections based on qualitative characteristics relevant for the modeler rather than probability or random sampling schemes, ultimately appointing members that are most independent from the remaining members of the ensemble.

2.6 Performance evaluation and computational methods

The biases between each member of the multimodel-ensemble and historical observations were assessed in terms of several metrics namely: the normalized root-mean square error (*nRMSE*), the mean bias error (*MBE*), the modified index of agreement (*MDA*), the percentage bias (*PBIAS*) and the mean absolute error (*MAE*). These objective functions have extensively been employed in meteorological and hydrological evaluation studies [21-23]. All data processing was executed using the R programming language [24]. Geostatistical modeling and spatio-temporal data analysis was implemented by combining functionalities of the *ncdf4*, *eurocordexr*, *raster* and *climdex* packages. PCA and HC analysis were generated by means of the *FactoMineR*, *factoextra* and *stats* packages, whereas objective functions were supplied by the *PerformanceAnalytics* and *ModelMetrics* R packages. An Intel® Xeon® Platinum 8160M Processor @ 3.70 GHz (24 cores, 48 threads) with 128 GB-RAM running GNU-Linux Debian 12 (*Bookworm*) was used to run experiments in parallel using R libraries *doParallel* and *foreach* whenener possible. Since the execution time of the cost-functions vary with each multimodel-ensemble member, the computational cost is herein presented in terms of user time instead of CPU time.

3 Results and discussion

3.1 Climatic precipitation indices

The CORDEX-CA 19-member multimodel-ensemble is driven by 10 different GCMs forced by 4 RCMs (Table 1). The HadGEM2, MPI, and GFDL GCMs appear three times using different RCMs and domains (CAM44 and CAM 22), whereas CanESM, CNRM and NorESM appear two times, with different driving models and spatial resolutions likewise. The remaining GCMs (CM5A, CSIRO, EARTH and MIROC) are all solely associated to the RCA regional climate model, but linked to a coarser horizontal resolution of 50 km (CAM44). As aforementioned, the ETCCDI indices were selected based on their performances in capturing specific properties of the precipitation patterns. Such properties are expressed within the statistic-relevance of the indices themselves. As expected, large discrepancies were found within ETCCDI precipitation indices among simulations of ensemble-members and between ensemble-members and observations from the SJO weather station. Indices of total precipitation volume (prcptotal, cwd and cdd) show distinctly large spreads, with considerably wetter conditions for all GCMs reduced by RegCM and CRCM, while the opposite behaviour is shown by GCMs reduced by REMO with extremely drier conditions, all of them at 25 km resolution (CAM22). Contrastingly, GCMs reduced by RCA at 50 km spatial resolution (CAM44) show noticeably larger spreads than those at 25 km resolution, with extreme wet and dry tendencies. GCMs such as HadGEM2, MPI and GFDL, all of them reduced by the four available RCMs, show diametrically opposite dry-wet tendencies for the reference period regardless of spatial resolution, which clearly demonstrates that RCMs add an additive layer of uncertainty due to random and systematic biases in the physics-structure, boundary conditions, and parameterization of each RCM [2].

Table 1. Summary of the average ETCCDI precipitation indices for CORDEX-CA simulated projections vs. observations from the SJO station for the reference period.

GCM_RM	Res.	prcp [mm]	cwd [d]	cdd [d]	r10 [d]	r20 [d]	r30 [d]	r50 [d]	r95p [mm]	r99p [mm]	rx1 [mm]	rx5 [mm]	sdii [mm]
SJO-Observations	**[-]**	**1882**	**15**	**64**	**63**	**31**	**16**	**4**	**398**	**112**	**139**	**238**	**14**
HadG-EM2_RegCM	0.22	5219	50	22	86	58	44	29	2125	757	971	1463	24
GFDL_RegCM	0.22	3461	63	18	88	50	33	14	1056	415	576	806	15
MPI_RegCM	0.22	5103	94	18	125	77	51	23	1550	642	678	1157	21
HadG-EM2_REMO	0.22	129	6	31	1	1	1	1	27	16	55	78	2
NorESM_REMO	0.22	90	4	59	2	1	1	1	35	14	87	91	3
MPI_REMO	0.22	192	6	32	3	2	1	1	83	43	203	311	3
CanESM_CRCM	0.22	1589	37	17	45	14	5	2	369	120	99	239	7
GFDL_CRCM	0.22	4668	85	11	111	62	44	25	1425	426	210	762	17
CNRM_CRCM	0.22	3099	40	11	66	34	24	12	1107	389	257	593	12
HadGEM2_RCA	0.44	4638	42	54	117	85	61	25	993	321	327	780	25
CanESM_RCA	0.44	564	7	92	16	8	4	2	183	64	182	358	10
EARTH_RCA	0.44	6185	75	31	157	124	88	32	1138	367	285	663	28
CM5A_RCA	0.44	426	6	132	12	6	4	2	150	67	266	528	10
NorESM_RCA	0.44	429	8	66	15	5	2	1	111	34	88	169	6
CNRM_RCA	0.44	5780	64	47	149	112	80	34	986	275	311	595	28
GFDL_RCA	0.44	4105	59	51	115	83	54	20	730	217	273	657	23
MIROC_RCA	0.44	3278	32	62	83	57	40	18	760	259	287	733	22
CSIRO_RCA	0.44	3017	26	58	67	46	33	18	869	283	382	824	22
MPI_RCA	0.44	7212	82	30	169	137	106	50	1182	351	287	579	32
HadG-EM2_RegCM	0.44	5219	50	22	86	58	44	29	2125	757	971	1463	24

Where: prcptotal: Annual total precipitation; cwd: consecutive wet days; cdd: consecutive dry days; r10: Annual count of days when precipitation ≥ 10mm; ; r20: Annual count of days when precipitation ≥ 20mm; ; r30: Annual count of days when precipitation ≥ 30mm; ; r50: Annual count of days when precipitation ≥ 50mm; r95p: sum of precipitation exceeding the 95th percentile of daily precipitation; r99p: sum of precipitation exceeding the 99th percentile of daily precipitation; rx1: 1-day precipitation; rx5: 5-day precipitation; sdii: simple precipitation intensity index. **Blue colors** donate wetter conditions while **red colors** indicate drier conditions in respect of historical observations.

Additionally, Regional Climate Models inherent systematic errors from GCMs simulations [3]. Indices of heavy precipitation days (r10, r20, r30 and r50) and of wet days (r95p, r95p and sdii) show evidently larger spreads than the other indices with significantly larger uncertainties of the simulation among combinations of GCMs and RCMs. However, there is a clear drier tendency for GCMs reduced by REMO, and the CanESM, CM5A and NorESM GCMs reduced by RCA4. By contrast, all remaining GCM-RCM combinations exhibit a distinctive wetter pattern. Regarding extreme precipitation indices (rx1 and rx5), there is a significant overestimation in the magnitude of the historical events in most cases, with GCM-RCMs (e.g. HadGEM2_RegCM, HadGEM2_RegCM, GFDL_CRCM, etc.) reaching over twice the observed values (139 and 238 mm respectively). Collectively, the average ensemble outlines a much wetter tendency when compared to historical observations. All things considered, the ETCCDI precipitation indices proved to adequately capture the main statistical features of the precipitation field associated to CORDEX-CA members linked to the SJO station in a compact and efficient way.

3.2 PCA Principal component analysis

Based on the ETCCDI precipitation indices, PCA analysis identified eleven PCs (Table 2). The first two components describe 88.81% of the variance in the initial variables, while the first three components describe 97.02% of the variance. If random variances are equal or larger than the historical observed ones, the corresponding PCs can be considered as noise and can be excluded [23]. Therefore, the dimensionality was reduced by retaining only the first three robust PCs. The PCA variable correlation plot for the first three PCs shows the structural relationship between variables and components (Figure 2). The Pearson correlation coefficient (ranging from -1 to 1) between a variable and a component can be read directly by projecting a variable vector onto the component axis, showing in which direction the variables correlate [15]. For instance, PC1 correlation is read from left to right, whereas PC2 correlation is read from bottom to top. Alternatively, the correlation coefficient is also presented in the form a heatmap (Figure 3 (a)). Regarding PC1 vs. PC2 (73.75% and 15.06% of the variance respectively), extreme precipitation indices (rx1 and rx5) and wet days indices (r95p and r95p) highlight a distinctive group (group-1) (Figure 2 (a)), whereas indices dealing with total precipitation volumes (prcptot and cwd) and heavy precipitation (r10, r20, r30 and r50) form another divergent group (group-2) both of which, present a strong positive correlation with PC1 (corr > 0.90) (Figure 3 (a)). Conversely, group-1 positively correlates with PC2, while group-2 negatively correlates with PC2, in both cases with considerably weaker correlations when compared to PC1, either positive or negative (Figure 3 (a)). The only clear exception in this case seems to be cdd (corr ~ -0.50), which negatively correlates with PC1. This is to be expected since cdd values deeply underestimates the average observed value (Table 1), indicating that in general, the ensemble is wetter than the average observed counterpart.

Table 2. Explained variance of principal components (PCA) calculated from ETCCDI precipitation indices for CORDEX-CA simulated historical projections.

Eigenvalues	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC1O
Variance	8.11	1.66	0.90	0.16	0.09	0.04	0.02	0.01	0.00	0.00
% of var.	73.75	15.06	8.21	1.49	0.83	0.38	0.19	0.07	0.02	0.01
Cum.%var.	73.75	88.81	97.02	98.51	99.34	99.72	99.90	99.97	99.99	100.00

The same inclination seems to apply for PC1 vs. PC3 (73.75% and 8.21% of the variance respectively), where most ETCCDI indices tend to form a unique group (group-1) aligned along PC1, with cdd once more being the only diverging variable (Figure 2 (b)). Moreover, correlations for PC3 are notably lower when compared to PC1 and PC3 (Figure 3 (a)). Conversely, PC1 and PC2 variable contributions are proportionally similar for most ETCCDI indices suggesting a fair representation of the variables on the PCs (Figure 3 (b)). Nonetheless, cdd contributions for PC1 and PC2 are substantially lower (~ 3.12 % and 0.96 % respectively) indicating that the variable is not properly represented by either PC. Notwithstanding, the opposite result is shown for PC3, where most of the contribution is in fact associated to cdd (~ 78.75 %) and the remaining indices have negligible contributions (except rx1 and rx5). In essence, cdd is the only divergent ETCCDI index that at the same time, is mostly contributing to PC3. This should not suggest that such a pattern would necessarily replicate at other locations. In summary, PCA proved competent in reducing the dimensionality by retaining only the first three robust PCs, describing 97.02% of the variance linked to precipitation ETCCDI indices [14]. At the same time, PCA was qualified to verify heavy interdependencies among GCM-RCM members and to confirm double-counting [16].

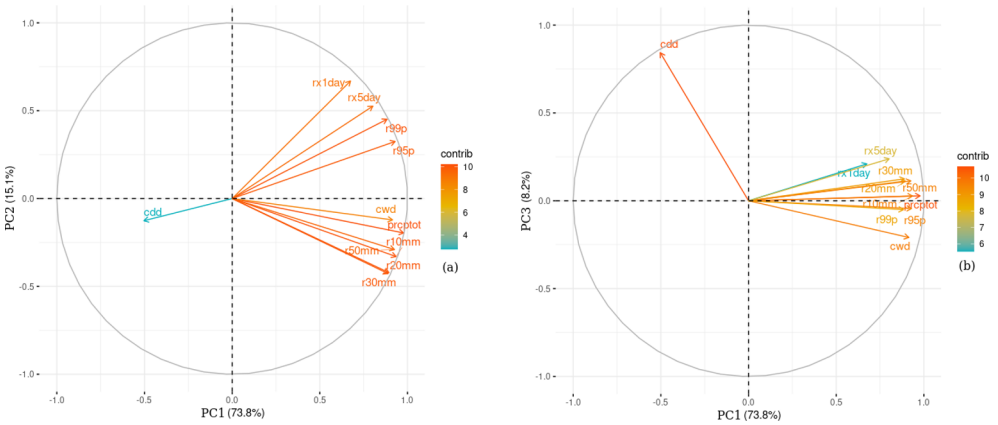

Fig. 2. PCA variable correlation plot for PC1 vs. PC2 (a) and PC1 vs. PC3 (b) for ETCCDI precipitation indices for CORDEX-CA simulated historical projections.

3.3 HC Hierarchical clustering

The broken-stick method applied to the Ward's criterion hierarchical clustering (CA) identified four closely related clusters of common attributes within the CORDEX-CA ensemble, which tree-like dependency structure is presented by the corresponding dendrogram (Figure 4). The GCMs forcing of the RCMs is a crucial component for model similarity in this context.

Notably, within cluster-1, simulations driven by lateral boundary conditions from NorESM and CanESM GCMs show strong specific and dense clustering, suggesting that RCMs (REMO, CRCM and RCA) powered by these particular GCMs perform rather similar in terms of precipitation patterns. Furthermore, NorESM and CanESM GCMs are only present in cluster-1 regardless of RCM or spatial resolution.

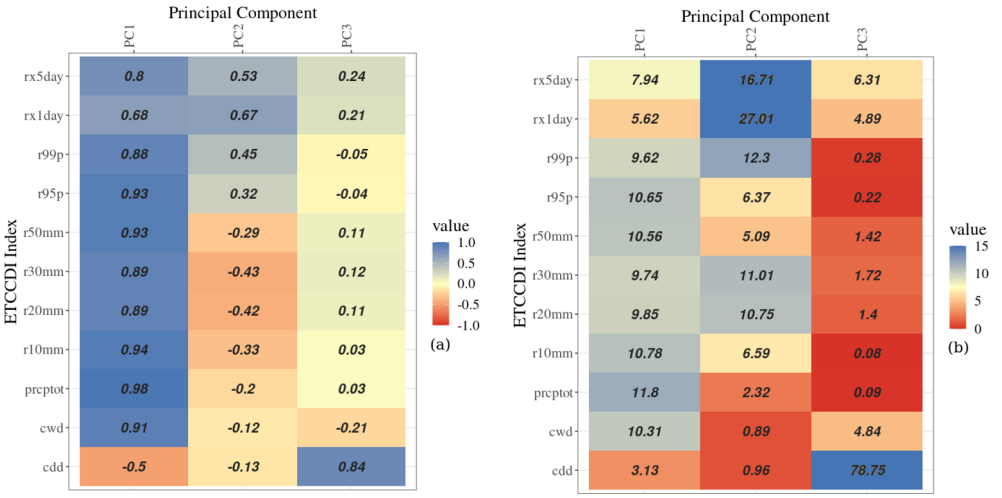

Fig. 3. Correlation (a) and contribution (b) PCA heatmaps for the first three components for ETCCDI precipitation indices for CORDEX-CA simulated historical projections.

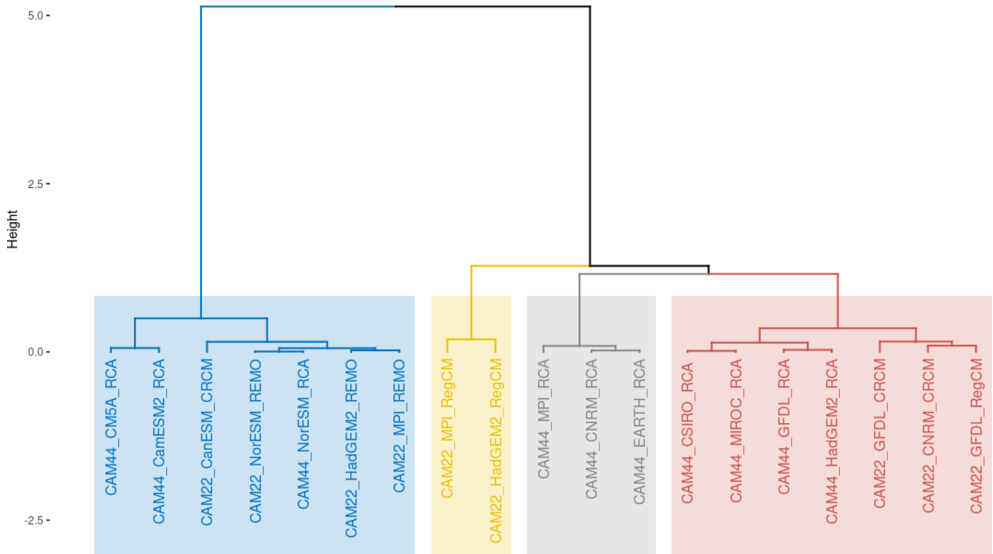

Fig. 4. Hierarchical clustering (HC) dendrogram based on ETCCDI precipitation indices for CORDEX-CA simulated historical projections.

This indicates that in many cases, RCMs with higher spatial resolutions and consequently more computationally intensive, do not necessarily translate into better representations of climatic mechanisms [25]. Hence, CAM22 should not be considered intrinsically better than CAM44. Interestingly, all simulations by REMO RCM are also part of cluster-1, where total-precipitation volume (prcptotal, cwd and cdd) and extreme precipitation indices (rx1 and rx5) reveal predominantly drier conditions along with CanESM_RCA, CM5A_RCA, NorESM_RCA and CanESM_CRCM. In principle, cluster-1 groups the driest members of the ensemble. Moreover, HadGEM2 and MPI GCMs, which also appear in cluster-1 (dry cluster), both reduced by RegCM, are the only two members comprising cluster-2, an extremely wet group exhibiting exceptionally high values of both total

precipitation volume (prcptotal, cwd and cdd) and extreme precipitation indices (rx1 and rx5) (Table 1). In contrast, cluster-3, integrated exclusively by RCA simulations driven by the lateral boundary conditions of the EARTH, CNRM and MPI GCMs displays even wetter precipitation volumes (prcptotal, cwd and cdd) when compared with cluster-2 but also significantly higher indices of heavy precipitation days (r10, r20, r30 and r50) that do not resemble any other ensemble member (Table 1). In addition, CNRM and MPI also exist in cluster-1 and cluster-2, displaying entirely opposite tendencies, since they are reduced by different RCMs. This confirms that GCMs dynamically reduced by different RCMs generate dramatically opposite outcomes, in this case explained only by profound differences in RCMs structures and parameterizations. It is also noteworthy that cluster-2 and cluster-3 are very heterogeneous in terms of GCMs, but inversely homogeneous regarding RCMs (RegCM and RCA respectively), splitting among different groups of similarity. Contrariwise, some driving GCMs do not create their own clusters at all (CM5A, CSIRO, EARTH and MIROC). Overall, cluster-2 represents a group of both extreme precipitation (rx1 and rx5) and wetter conditions (prcptotal, cwd and cdd), while cluster-3 stands for even wetter conditions (prcptotal, cwd and cdd) but heavier daily precipitation (r10, r20, r30 and r50). Meanwhile cluster-4, comparable in proportions to cluster-1, comprises members with surprisingly closer-to-observed heavy precipitation indices (r10, r20, r30 and r50) compared to the remaining models, indistinctively of RCM or spatial resolution even when both extreme precipitation (rx1 and rx5) and total precipitation volume indices (prcptotal, cwd and cdd) are indeed appreciably wetter than historical observations (Table 1). Accordingly, cluster-4, is one group that satisfactorily describes heavy precipitation storms but poorly characterizes all other ETCCDI indices. On the grounds of these results, it is clear that the combination of ETCCDI indices, PCA analysis and CA clustering is capable of yielding compact clusters of CORDEX-CA members that reasonably describes the primal statistical features of precipitation-patterns for the SJO weather station in a compact and efficient form [19].

3.4 Model similarity sampling

For visualization purposes, the four CORDEX-CA clusters identified by HC are featured in the principal component space for the first three PCs (Figure 5). The sign and order of magnitude of each ensemble-member corresponds to the pattern described in the corresponding PC from the components plot (Figure 2). Simulations with close-to-zero values can be interpreted as representing a precipitation average-pattern induced by the corresponding principal component [16]. As previously mentioned, the quota-sampling method was used to select one simulation out of each cluster of related models obtained from the HC analysis. Model dependency is significantly decreased by choosing one simulation from each cluster, yielding a more independent sub-ensemble as a result. For instance, an average-pattern simulation and 4 extreme representative simulations were chosen to span the uncertainty range and to obtain maximum diversity of the multimodel-ensemble subset. Additionally, the HC clusters show no dependencies in the PC1 vs. PC2 spectrum (Figure 5(a)) and only a slight dependency between cluster-3 and cluster-4 in the PC1 vs. PC3 space can be observed, which is considerably weaker than those of individual members (Figure 5(b)). Accordingly, the following five members were selected: CAM22_NorESM_REMO, CAM44_MIROC_RCA, CAM22_GFDL_RegCM, CAM44_MPI_RCA and CAM22_HadGEM2_RegCM, marked in the principal component space plot (Figure 5). Here, RegCM and RCA RCMs appear twice, but their corresponding GCMs trigger substantially different precipitation responses as previously stated [26]. This leads to a sub-ensemble where two RCMs appear twice while others are omitted entirely

(CRCM). At the same time, it is reasonable to assume that other combinations of models capturing extreme characteristics are also possible.

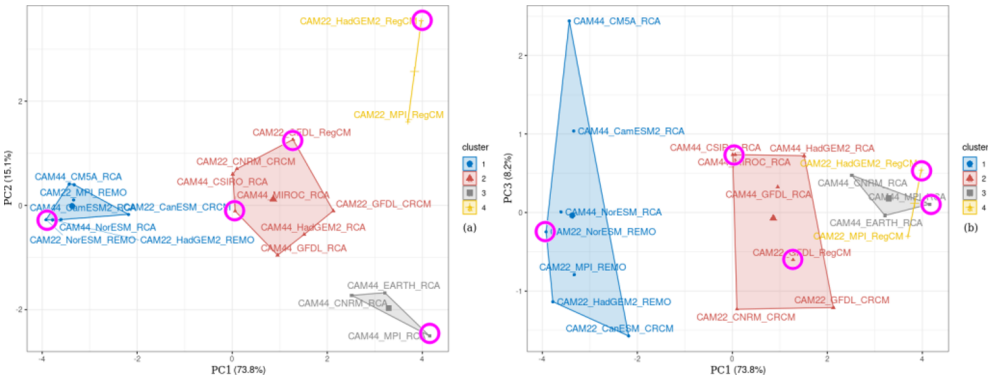

Fig. 5. HC clusters within the principal component space for the first three PCs based on ETCCDI precipitation indices for CORDEX-CA simulated historical projections. Subset selected members are marked by magenta circles.

3.5 Performance evaluation of model ensembles

The performance of the two multi-model ensemble-means, one using the entire 19-member ensemble (MEM-19) and the other using the 5-member sub-ensemble (MEM-5), was evaluated against the SJO station observational dataset for the reference period (Table 3). Regarding the selected objective functions, the sub-ensemble maintains the distinctive characteristics of the precipitation patterns of the entire ensemble. The RMSE, a measure of accuracy between observed and modeled values, which ranges from 0 to $+\infty$, presents slightly higher values for MEM-5, where closer to 0 values indicate a higher accuracy in estimation.

Table 3. Performance evaluation of multi-model ensemble-means (MEMs) for the CORDEX-CA precipitation historical projections.

MEM	nRMSE [mm/d]	MBE [mm/d]	MDA [%]	PBIAS [%]	MAE [mm/d]	Ave. Comput. Cost [sec]
MEM-19	15.53	3.14	0.42	56.97	10.47	2485
MEM-5	18.54	5.16	0.39	52.12	11.96	650

Similarly the MBE, which represents the average deviations between the two datasets, features marginally larger values for MEM-5. Positive MBE values indicate overestimation, while negative values indicate underestimation. Comparatively, the PBIAS, which calculates the relative volume difference between modeled and observed precipitation, shows remarkably elevated values of over 50% for both MEMs, with moderately smaller values for MEM-5. As previously stated, the ensemble depicts much wetter conditions than those historically observed, which is also confirmed by closely similar but at the same time high MAE positive values. Furthermore, the MDA, a robust modification of the Willmott index intended as a dimensionless measurement of model accuracy bounded by 0 meaning no agreement, and 1 meaning a perfect fit, shows undeniably deficient values of around

0.40 for either MEM, which suggests that in general terms, most GCM-RCM combinations cannot properly reproduce the highly convective precipitation patterns in altitude-dependence high-mountainous terrains with steep-slopes such as those characteristics of Costa Rica. Once more, this is the result of systematic biases caused by imperfect conceptualization and parameterization that should be resolved through the application of bias correction (BC) methods. Bias correction is the adjustment of biased simulated data to observations. BC methods aim to add value to model outputs by removing systematic biases of simulated data so that they can be used in climate change impact modeling [27-28]. As a consequence, post-processing of CORDEX-CA GCM-RCMs simulations through the application of bias correction (BC) methods seems unavoidable at least for the SJO weather station, irrespectively of MEM-19 or MEM-5, which altogether goes beyond the scope of this work. Irrespectively of the detected biases, no significant differences were found between the two MEMs by mean of the selected metrics, implying that the applied techniques are effective in reducing dimensionality, prevent double-counting of highly dependent simulations and considerably reduce the associated computational costs, as run times for MEM-5 are notably lower when compared to MEM-19.

4 Conclusions and recommendations

In this study, an approach to reduce the dimensionality of a multi-model ensemble (MME) built upon precipitation climate simulations from the Coordinated Regional Climate Downscaling Experiment (CORDEX) for the Central America domain (CA) based on the utilization of the Expert Team on Climate Change Detection and Indices (ETCCDI) was presented. The following main conclusions can be drawn: (a) the ETCCDI precipitation indices proved to adequately capture the main statistical features of the precipitation field associated to CORDEX-CA members linked to the SJO station in a compact and efficient way, (b) PCA was able to reduce the dimensionality by retaining only the first three robust PCs, describing 97.02% of the variance linked to precipitation ETCCDI indices. At the same time, PCA was capable of verifying heavy interdependencies among GCM-RCM members and to confirm double-counting, (c) HC clustering based on the Ward's criterion (CA) identified four closely related clusters of common attributes within the CORDEX-CA ensemble, (d) the broken-stick method significantly decreased model interdependencies and lowered random noise by choosing one simulation from each HC cluster, yielding a more independent sub-ensemble (MEM-5) as a result. Additional combinations of models capturing extreme characteristics are also possible, (e) post-processing of CORDEX-CA GCM-RCMs simulations through the application of bias correction (BC) is necessary for the SJO weather station regardless of MEM-19 or MEM-5 and (f) no significant differences were found between MEM-19 and MEM-5 by means of the selected metrics, implying that the applied techniques are effective in reducing dimensionality, prevent double-counting of highly dependent simulations and considerably reduce the associated computational costs. Likewise, the proposed methodology could be further improved by: (a) analysing additional HC combinations of ensemble-members, (b) including indices linked to other climatic indices (temperature, relative humidity), (c) incorporating the results of robust bias correction (BC) methods, (d) comparing with other linear and nonlinear dimensionality reduction techniques and (e) considering additional weather stations located across the various climatic regions of Costa Rica.

References

1. Taylor, K. E., Stouffer, R. J., & Meehl, G. A. (2012). An Overview of CMIP5 and the Experiment Design. Bulletin of the American Meteorological Society, 93(4), 485–498. https://doi.org/10.1175/BAMS-D-11-00094.1
2. Piani, C., Haerter, J. O., & Coppola, E. (2010). Statistical bias correction for daily precipitation in regional climate models over Europe. Theoretical and Applied Climatology, 99(1–2), 187–192. https://doi.org/10.1007/s00704-009-0134-9
3. Sharma, Ojha, Shukla, Pham, Linh, Fai, Loc, & Dung. (2019). Modified Approach to Reduce GCM Bias in Downscaled Precipitation: A Study in Ganga River Basin. Water, 11(10), 2097. https://doi.org/10.3390/w11102097
4. Masson, D., & Knutti, R. (2011). Climate model genealogy. Geophysical Research Letters, 38(8). https://doi.org/10.1029/2011GL046864
5. Mendez, M., Calvo-Valverde, L.-A., Imbach, P., Maathuis, B., Hein-Grigg, D., Hidalgo-Madriz, J.-A., & Alvarado-Gamboa, L.-F. (2022). Hydrological Response of Tropical Catchments to Climate Change as Modeled by the GR2M Model: A Case Study in Costa Rica. Sustainability, 14(24), 16938. https://doi.org/10.3390/su142416938
6. Gangrade, S., Kao, S. C. & McManamay, R. A. 2020 Multi-model hydroclimate projections for the Alabama-Coosa-Tallapoosa River Basin in the southeastern United States. Scientific Reports 10, 1–12. https://doi.org/10.1038/s41598-020-59806-6
7. Mendlik, T., & Gobiet, A. (2016). Selecting climate simulations for impact studies based on multivariate patterns of climate change. Climatic Change, 135(3–4), 381–393. https://doi.org/10.1007/s10584-015-1582-0
8. Sørland, S. L., Schär, C., Lüthi, D., & Kjellström, E. (2018). Bias patterns and climate change signals in GCM-RCM model chains. Environmental Research Letters, 13(7), 074017. https://doi.org/10.1088/1748-9326/aacc77
9. Bishop, C. H., & Abramowitz, G. (2013). Climate model dependence and the replicate Earth paradigm. Climate Dynamics, 41(3), 885–900. https://doi.org/10.1007/s00382-012-1610-y
10. Cannon, A. J. (2015). Selecting GCM Scenarios that Span the Range of Changes in a Multimodel Ensemble: Application to CMIP5 Climate Extremes Indices. Journal of Climate, 28(3), 1260-1267. https://doi.org/10.1175/JCLI-D-14-00636.1
11. Bethere, L., Sennikovs, J., & Bethers, U. (2017). Climate indices for the Baltic states from principal component analysis. Earth System Dynamics, 8(4), 951–962. https://doi.org/10.5194/esd-8-951-2017
12. Waylen, P. R., Quesada, M. E., & Caviedes, C. N. (1996). Temporal and spatial variability of annual precipitation in Costa rRica and the southern oscillation. International Journal of Climatology, 16(2), 173–193. https://doi.org/10.1002/(SICI)1097-0088(199602)16:2<173::AID-JOC12>3.0.CO;2-R
13. Mendez, M., Calvo-Valverde, L.-A., Hidalgo-Madriz, J.-A., & Araya-Obando, J.-A. (2023). A comparison of generalized extreme value, gumbel, and log-pearson distributions. A case study in Costa Rica. BIO Web Conf., 62 (2023) 01002. https://doi.org/10.1051/bioconf/20236201002
14. Leander, R., Buishand, T. A., & Tank, A. M. G. K. (2014). An Alternative Index for the Contribution of Precipitation on Very Wet Days to the Total Precipitation. Journal of Climate, 27(4), 1365–1378. https://doi.org/10.1175/JCLI-D-13-00144.1

15. Kraemer, G., Reichstein, M., & Mahecha, M., D. (2018). dimRed and coRanking—Unifying Dimensionality Reduction in R. The R Journal, 10(1), 342. https://doi.org/10.32614/RJ-2018-039
16. Benestad, R., Parding, K., Dobler, A., & Mezghani, A. (2017). A strategy to effectively make use of large volumes of climate data for climate change adaptation. Climate Services, 6, 48–54. https://doi.org/10.1016/j.cliser.2017.06.013
17. Peres, D. J., Senatore, A., Nanni, P., Cancelliere, A., Mendicino, G., & Bonaccorso, B. (2020). Evaluation of EURO-CORDEX (Coordinated Regional Climate Downscaling Experiment for the Euro-Mediterranean area) historical simulations by high-quality observational datasets in southern Italy: Insights on drought assessment. Natural Hazards and Earth System Sciences, 20(11), 3057–3082. https://doi.org/10.5194/nhess-20-3057-2020
18. Xu, R., Chen, N., Chen, Y., & Chen, Z. (2020). Downscaling and Projection of Multi-CMIP5 Precipitation Using Machine Learning Methods in the Upper Han River Basin. Advances in Meteorology, 2020, 1–17. https://doi.org/10.1155/2020/8680436
19. Singh, S. K., Lo, E. Y.-M., & Qin, X. (2017). Cluster Analysis of Monthly Precipitation over the Western Maritime Continent under Climate Change. Climate, 5(4), 84. https://doi.org/10.3390/cli5040084
20. Murtagh, F., & Legendre, P. (2014). Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? Journal of Classification, 31(3), 274–295. https://doi.org/10.1007/s00357-014-9161-z
21. Mendez, M., Calvo-Valverde, L.-A. (2016). Development of the HBV-TEC Hydrological Model. Procedia Engineering, 154, 1116-1123. https://doi.org/10.1016/j.proeng.2016.07.521
22. Reiter, P.; Gutjahr, O.; Schefczyk, L.; Heinemann GCasper, M. Does applying quantile mapping to subsamples improve the bias correction of daily precipitation? Int. J. Climatol. 2018, 38, 1623–1633. https://doi.org/10.1002/joc.5283
23. Jose, D. M., Vincent, A. M., & Dwarakish, G. S. (2022). Improving multiple model ensemble predictions of daily precipitation and temperature through machine learning. Scientific Reports, 12(1), 4678. https://doi.org/10.1038/s41598-022-08786-w
24. R Core Team (2023). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/
25. Almazroui, M., Islam, M. N., Saeed, F., Saeed, S. (2021). Projected Changes in Temperature and Precipitation Over the United States, Central America, and the Caribbean in CMIP6 GCMs. Earth Systems and Environment, 5(1), 1–24. https://doi.org/10.1007/s41748-021-00199-5
26. Oyerinde, G. T., Lawin, A. E., & Anthony, T. (2022). Multiscale assessments of hydroclimatic modelling uncertainties under a changing climate. Journal of Water and Climate Change, 13(3), 1534–1547. https://doi.org/10.2166/wcc.2022.266
27. Mendez, M., Calvo-Valverde, L.-A. (2020). Comparison performance of machine learning and geostatistical methods for the interpolation of monthly air temperature over Costa Rica. IOP Conf. Ser.: Earth Environ. Sci., 432, 012011. https://doi.org/10.1088/1755-1315/432/1/012011
28. Pereira, H. R., Meschiatti, M. C., Pires, R. C. D. M., & Blain, G. C. (2018). On the performance of three indices of agreement: An easy-to-use r-code for calculating the Willmott indices. Bragantia, 77(2), 394–403. https://doi.org/10.1590/1678-4499.2017054