



Downscaled CMIP6 climate model subset selection

C.L. Curry, D. Ouali, S.R. Sobie and F.W. Zwiers Regional Climate Impacts group, Pacific Climate Impacts Consortium 18 July 2023





This report outlines a method for selecting a subset of earth system models (ESMs) from the Sixth Coupled Model Intercomparison Project (CMIP6) that is sufficiently representative of an ensemble of 26 models from CMIP6 for Canada and its subregions. The specific objective is to obtain a subset of reasonably independent ESMs that captures the overall range of projected change in a representative set of climate extremes (ETCCDI or Climdex) indices constructed from the ESM outputs. Projections are calculated for a future epoch corresponding to a global mean temperature change of 2 °C relative to 1971-2000, using results from two of the CMIP6 Shared Socioeconomic Pathways (SSPs), SSP2-4.5, and SSP5-8.5. The selection procedure is described below and representative subsets are provided for Canada and five of its subregions.

1. Initial application and comparison of subset selection algorithms

Representative subsets of the CMIP5 ensemble of ESMs were previously determined using selection criteria (Cannon, 2015) designed to capture the range of projected changes from the Climdex indices of extremes (Zhang et al., 2011). Models were chosen iteratively using the objective Katsavounidis–Kuo–Zhang (KKZ) algorithm (Katsavounidis et al., 1994), to provide a subset of ESMs that spanned at least 90% of the range of standardized, future-projected changes in each of the 27 Climdex indices for the full CMIP5 ensemble¹. The CMIP5 subset of 12 ESMs selected using the KKZ method for Western North America (commonly referred to as "PCIC12") has been used for a variety of research and climate impacts applications.

The selection procedure used previously for CMIP5 has some disadvantages, however. First, there is a certain degree of degeneracy in the procedure in two respects. Many ESMs in the full ensemble are mutually dependent due to their "shared history" of development (Brunner et al., 2020; Section 2), and certain Climdex indices are more or less redundant as they characterize similar phenomena (Section 3). Second, raw ESM outputs were used in the CMIP5 procedure, even though the intention was to produce subsets of downscaled ESMs. Third, the regional definitions used for CMIP5 (Giorgi Regions), which spanned North America while mixing parts of Canada and the U.S., are not particularly useful for regional studies within Canada alone. Fourth, the scenario-dependence of the selection procedure was not previously explored, and different subsets may be more appropriate for different future emissions scenarios. As described in Section 4 below, we address this by combining model results across scenarios at a specified level of global warming.

Finally, the KKZ method lacks an explicit constraint to maintain the ensemble median or average of the full ensemble within the selected subset. Since it selects outlier models by design, the method preferentially selects ESMs with strong, rather than weak, responses to forcing. In the context of the CMIP6 ensemble, whose temperature distribution displays a strong positive skewness within a specified future time period (Sobie et al., 2021), this means that the method tends to select "hot models" (Hausfather et al., 2022), thereby raising the

¹ In practice, a generalized distance matrix is constructed between the index changes in the subset and the corresponding changes in the full ensemble, normalized by the standard deviation of each index; see Cannon (2015).





subset mean and median compared to that of the full ensemble. However, by applying the cross-scenario selection method in combination with KKZ, we have been able to mitigate this tendency. We describe each of these improvements more fully in the sections which follow.

2. Gauging model independence

While the CMIP6 ensemble comprises roughly 30 ESMs from climate modelling centres worldwide, it is recognized that this is not a set of independent models. Several of the models have identical or similar components, which introduces some level of degeneracy amongst them and suggests that subsetting might begin with the removal of one or more ESMs from the same "family tree." We use two resources to carry out this initial step, one qualitative and one quantitative. First, information on each ESM's components is taken from Annex II of the recent IPCC AR6 report (IPCC, 2021). Second, we consulted Brunner et al. (2020) (hereafter B20) who constructed a model-model generalized distance matrix for 33 CMIP6 ESMs based on climatological global mean temperature and sea level pressure over the period from 1980-2014. B20 noted that models with shared components or with the same origin (according to the preceding document) are always grouped together in the colour-coded dendrogram computed from the inter-model distance matrix (see their Figure 5). B20 also calculated an "independence shape parameter," σ_s , in units of the same generalized distance, that suggests the removal of 10 ESMs determined to be closely related.

Of the 26 CMIP6 ESMs downscaled at PCIC (see Table 1), 22 were included in B20's dendrogram. Of these, we selected at most two ESMs from each colour group separated by a generalized distance $\sigma_s > 0.54$, to obtain a subset of 13 nominally independent ESMs. Four ESMs that we downscaled were not considered by B20: TaiESM1, KIOST-ESM, CMCC-ESM2 and NorESM2-LM. Of these, the KIOST-ESM was removed from consideration, given that it was built from similar components to GFDL-ESM4. This leaves the 16 nominally independent ESMs highlighted in **Table 1**. It is this reduced set, not the entire set of downscaled ESMs, that was used in the KKZ method in Section 4 below. Hereafter, this is referred to as the "reduced set."²

² During the subset selection process, we became aware of ~200 anomalously large maximum daily temperature events (some exceeding 60 °C) found in the raw output of the UKESM1-0-LL model at many Canadian grid cells. These originate from a problem that occasionally occurs at the boundary between the atmosphere and the JULES land surface scheme used in this ESM, as described in the Earth System Documentation Errata at errata.es-doc.org under:

[&]quot;Hadley Models: 76b3f818-d65f-c76b-bfd8-cae5bc27825c." For CanDCS-M6, we applied a correction scheme prior to the MBCn downscaling to remove these aberrant temperatures from this ESM. This correction was not implemented for the BCCAQv2-downscaled version of UKESM1, but is not expected to have a noticeable impact on the subset selection.





Table 1. Downscaled CMIP6 ESMs for Canada, initial selection of 16 ESMs based on model independence (shading),and timing of a global warming level (GWL) of 2°C for each model (central year of a 31-year moving window)relative to the 1971-2000 global mean temperature.

	Deumonale d 50M	Timing of GWL = 2°C			
	Downscaled ESIVI	SSP1-2.6	SSP2-4.5	SSP5-8.5	
1	ACCESS-CM2	2053	2047	2042	
2	ACCESS-ESM1-5	2084	2054	2044	
3	BCC-CSM2-MR	-	2077	2050	
4	CanESM5	2048	2037	2032	
5	CMCC-ESM2	2068	2055	2049	
6	CNRM-CM6-1	2084	2063	2049	
7	CNRM-ESM2-1	2085	2065	2051	
8	EC-Earth3	-	2065	2051	
9	EC-Earth3-Veg	2085	2062	2047	
10	FGOALS-g3	-	2085	2058	
11	GFDL-ESM4	-	2085	2059	
12	HadGEM3-GC31-LL	2045	2037	2032	
13	INM-CM4-8	-	2085	2057	
14	INM-CM5-0	-	2085	2058	
15	IPSL-CM6A-LR	2073	2053	2046	
16	KACE-1-0-G	2057	2043	2037	
17	KIOST-ESM	_	2083	2057	
18	MIROC6	_	2085	2060	
19	MIROC-ES2L	-	2082	2054	
20	MPI-ESM1-2-HR	-	2085	2062	
21	MPI-ESM1-2-LR	-	2085	2062	
22	MRI-ESM2-0	2065	2063	2047	
23	NorESM2-LM	-	2085	2059	
24	NorESM2-MM	_	2085	2058	
25	TaiESM1	2041	2042	2035	
26	UKESM1-0-LI	2038	2035	2032	





3. Climate index selection

PCIC's CMIP5 ESM selection procedure used the Climdex set of climate extreme indices (Zhang et al., 2011; see Figure 1 and www.climdex.org). These indices are based on daily maximum and minimum temperature and daily total precipitation. There is naturally some redundancy over the set: e.g., Frost Days (FD) vs. Ice Days (ID), temperature extremes (TXx vs. TX90p, TNn vs. TN10p). A few indices are not in fact extremes: e.g., total precipitation (PRCPTOT), simple daily precipitation intensity (SDII), and growing season length (GSL). While applying the KKZ algorithm, it was noticed that the range of certain Climdex indices spanned by the subset was significantly and consistently poorer than others, which led to the aforementioned increase in the subset size. These indices, which are all threshold-based, tend to exhibit very small changes from historical values in future periods (see www.climdex.org for definitions): CWD, CDD, R1mm, TX10p, TN10p, TR, CSDI and DTR. Problems with the simulation of trace precipitation (which underlie CWD, CDD and R1mm) by most ESMs have been well documented, and contribute to these indices being less reliable. Indices probing the lower tail of the frequency distribution of temperature (TX10p, TN10p, and CSDI), are subject to the shrinking area of the lower tail under future warming. By definition, the first two indices have an effective upper bound of 10% in the reference period, approaching zero as the entire temperature distribution shifts toward higher values in future (with the opposite being true of TN90p and TX90p). Model spread is therefore artificially constrained to the 10% value in the far future for these variables, making it a non-robust target for KKZ (the reasoning is conceptually the same for CSDI, based on the same threshold as TN10p). Large inter-daily variability in the daily temperature range from which DTR is constructed produces small projected differences for this index that are not robust across different ESMs. Finally, the high nighttime temperature threshold of TR (20°C) is rarely exceeded over most of Canada, even in future, making it a noisy and unsuitable metric for discriminating amongst models.

For these reasons, we omitted the eight Climdex indices identified above from the subset selection using the KKZ method. Two additional changes to the index set were also made. First, the index SDII was removed since it provides essentially the same information as PRCPTOT. Second, due to its importance for constraining the median temperature response, we added annual mean temperature, TANN, to the set of indices. The final set of 19 indices used to evaluate model subsets, comprising 12 based on daily temperature and 7 based on daily precipitation, is contained in **Table 2**.

Heat and cold					Precipitation				
FD	SU	ID	GSL	WSDI	TXx	Rx1day	Rx5day	R10mm	R20mm
TNn	TNx	TXn	TN90p	TX90p	TANN	R95p	R99p	PRCPTOT	

 Table 2. Subset of Climdex indices used in this study.





4. Cross-scenario selection strategy to obtain representative subsets

For CMIP5, representative subsets of the CMIP5 ESM ensemble were determined using the KKZ procedure applied to raw model differences between the 2071-2100 and 1971-2000 periods. The ESM projections were taken from the RCP4.5 scenario, and subsets determined for a number of geographic sub-regions of North America (Pacific Climate Impacts Consortium, 2019). Different subsets were obtained for other RCPs, but for simplicity only the RCP4.5 subsets were recommended for use. Moreover, it was assumed that the subsets would be equally valid for the downscaled and bias-corrected results (now known as the CanDCS-U5 dataset). Here, we improve on the previous procedure in both respects, namely by: 1) using climate indices calculated from the downscaled and bias-corrected CMIP6 ESMs, and; 2) using results from more than one future emissions scenario in the KKZ method, to produce a single representative subset for each region. Also, for CMIP6, two distinct downscaling methods were applied to the 26 ESM set: BCCAQv2, which produced the CanDCS-U6 dataset, and MBCn, which produced the CanDCS-M6 dataset. We provide representative subsets for both downscaled sets of ESMs.

An important characteristic of ESM projections, demonstrated in numerous studies (e.g., Seneviratne et al., 2016), is that many regional or local climate indicators change roughly in proportion to a specified rise in global mean temperature, *independent of emissions scenario*. This presents an opportunity to derive ESM subsets that are applicable under a range of possible future conditions, and therefore suited to a host of applications. Choosing a global mean warming level (GWL) of 2°C (relative to the 1971-2000 global mean temperature) for specificity, we determined the time of exceedance—i.e., the central year of a 31-year moving average at which this GWL is irrevocably exceeded—for each of ESM-SSP combination. The results, displayed in the right-hand columns of Table 1, show that while all of the reduced set of 16 ESMs reach the 2°C threshold for SSP5-8.5 and SSP2-4.5, only 9 ESMs reach this value under SSP1-2.6.³ Therefore, to avoid an unequal representation amongst simulations from different scenarios, we omit the SSP1-2.6 results in the cross-scenario selection procedure, which proceeds as follows:

1. For each ESM-SSP combination in the reduced subset, calculate the change in each climate index compared to its 1971-2000 baseline value, at the time of that ESM-SSP reaching the 2°C level. For example, for the CNRM-ESM2-1 forced by SSP5-8.5, the future period is 2035-2066, centred on the year 2051.

2. For each ESM, obtain projected index changes from both SSPs for each of the 19 indices considered, yielding a vector of 38 projected index changes at the 2°C GWL.

³ These results are drawn from Appendix A of the paper by Sobie, Zwiers & Curry (2021). Although the timings reported there (and in Table 1) were computed from raw ESM results, we checked that the differences in annual mean temperature and precipitation anomalies using the downscaled ESM results for the same periods were negligible.





3. Using these 38 index changes as inputs, construct an ordered list of ESMs according to the KKZ method (Section 1).

4. Select the final subset of ESMs from this list by requiring that the 10th percentile of the distribution of index changes exceeds 85% of the range of projections as calculated from the reduced set of ESMs (as illustrated below).

Figure 1 (left) shows how the range of index changes, relative to the corresponding range in the 16-member set, evolves as ESMs are added to the subset by the KKZ algorithm. Once the 12th ESM is added, the lower whisker indicating the 10th percentile of results for all 19 indices rises above 80%, terminating the selection process. While several indices still exhibit poorer coverage, this behaviour doesn't change as additional models are added (until the 15th and 16th ESMs added, when, by construction, the covered range is 100% for all indices).



Figure 1. *Left:* Standard boxplots showing the covered range (in percent) of standardized, projected change (at a GWL = 2°C relative to 1971-2000) in 19 Climdex indices, captured by the selected subsets using the KKZ selection method applied over the entire Canadian landmass. The red segments indicate the 10th percentile of index changes. *Right:* Scatter plot of change in annual mean temperature against change in annual total precipitation, over all of Canada, for the full CanDCS-U6 ensemble (*grey symbols*) and for the reduced ensemble of 16 ESMs (*purple symbols*). The grey and purple plus signs indicate the corresponding means of the two ensembles. The KKZ-selected subsets are displayed with symbols numbered 1 through 12, with the green numbered symbols showing values for SSP2-4.5, and the blue numbered symbols showing values for SSP5-8.5. The KKZ method selects five ESMs located on the convex hull enclosing the points, and seven ESMs in the interior of the hull. The boxplots along the sides of the panel indicate the distribution of changes in TANN (*right*) and PRCPTOT (*top*) for the full (*grey*) and reduced (*purple*) ensembles and the KKZ-selected subset (*red*).

Since the KKZ method employs a generalized distance matrix in a 2*N*-dimensional space (where *N* is the number of climate indices), the results for specific climate indices can be difficult to visualize. Nevertheless, **Figure 1** (left) illustrates one outcome of the procedure, i.e., a scatterplot of changes in just two key indices, annual mean temperature and annual total precipitation. The plot shows how changes from the KKZ-selected ESMs (numbered symbols)





compare to those from both the full and reduced CMIP6 ensembles. It can be seen that several of the selected ESMs define the "convex hull" of bivariate extremes from the 16-member reduced set, with the remaining selected models occupying the interior of the hull. The boxplots along the outside of the plot compare the distributions of TANN and PRCPTOT from the KKZ-selected subset for Canada with those from the full and reduced CMIP6 ensembles. The KKZ subset spans the full ranges of the reduced ensemble, with median values only slightly higher for both variables, by about +0.1°C for TANN and +0.5% for PRCPTOT.

The final subsets for Canada are listed in the tables below, and comprise the same 12 ESMs for CanDCS-U6 (**Table 3**) and CanDCS-M6 (**Table 4**). It is reassuring that the cross-scenario selection method results in national-scale ESM subsets that are insensitive to the downscaling method.⁴

5. Application of the subset selection strategy regionally within Canada

Given that some users of climate projections are interested primarily in results at a regional and/or local scales, we applied the same selection strategy described in Section 4 to five large regions of Canada, as depicted in **Figure 2**.



Figure 2. Regions of Canada used for downscaled ESM subset selection.

⁴ It is evident from the tables that although the Canada-wide subsets contain the same ESMs, they are selected in slightly different order by the KKZ scheme. This simply reflects how the spread in generalized distance between the subset and the reduced ensemble evolves as the method proceeds (Section 2). Hence, other than the selection of the first-ranked model, which is closest to the ensemble mean, *the ordering in Tables 3 and 4 should not be interpreted as a quality-based performance ranking amongst the ESMs*.





All 26 GCMs 16 GCMs KKZ ssp245 ssp585

20

15

10

Change in PRCPTOT (%)

8 0 0

Distinct climates between the regions lead to corresponding differences in projections, meaning that a subset that represents Canada well as a whole may not be equally representative of a region (by our specific definition of "representativeness"). Users interested in climate projections for a single region are advised to use the corresponding regional subset listed in Tables 3 and 4, while those studying more than one region should use the Canada-wide subsets.

Figure 3 shows results equivalent to Figure 1, but for the BC region, with similar results seen for the other subregions. The number of ESMs in the regional subsets ranges from 8 to 11, depending on the downscaling method and region. As for the national subsets, there is a high degree of overlap between the downscaled ESMs from the two methods.

40

3.5

3.0

2.5

2.0

5

Change in TANN (°C)



Figure 3. Same as Figure 1, but for the BC region.





Table 3. Cross-scenario, KKZ-selected ESM subsets for Canada and each of its subregions, from the reduced set of

 CanDCS-U6 downscaled CMIP6 scenarios. Models in italics differ between the CanDCS-U6 and CanDCS-M6 subsets.

	CANADA	Eastern	Ontario	Prairies	British	Northern
		Canada			Columbia	Canada
1	BCC-CSM2-	GFDL-ESM4	BCC-CSM2-	BCC-CSM2-	TaiESM1	CanESM5
	MR		MR	MR		
2	NorESM2-	NorESM2-	NorESM2-	EC-Earth3-	NorESM2-	INM-CM5-0
	LM	LM	LM	Veg	LM	
3	MIROC-ES2L	MIROC-ES2L	MIROC-ES2L	UKESM1-0-	CNRM-	NorESM2-
				LL	ESM2-1	LM
4	MPI-ESM1-	FGOALS-g3	UKESM1-0-	NorESM2-	IPSL-CM6A-	MRI-ESM2-0
	2-HR		LL	LM	LR	
5	MRI-ESM2-0	MRI-ESM2-0	EC-Earth3-	FGOALS-g3	MIROC-ES2L	MPI-ESM1-
			Veg			2-HR
6	UKESM1-0-	EC-Earth3-	ACCESS-	INM-CM5-0	MRI-ESM2-0	ACCESS-
	LL	Veg	ESM1-5			ESM1-5
7	EC-Earth3-	INM-CM5-0	INM-CM5-0	MRI-ESM2-0	UKESM1-0-	CMCC-ESM2
	Veg				LL	
8	CMCC-ESM2	CanESM5	GFDL-ESM4	MPI-ESM1-	EC-Earth3-	UKESM1-0-
				2-HR	Veg	LL
9	INM-CM5-0		CNRM-	CNRM-	MPI-ESM1-	
			ESM2-1	ESM2-1	2-HR	
10	FGOALS-g3		CanESM5	IPSL-CM6A-	FGOALS-g3	
				LR		
11	TaiESM1		MRI-ESM2-0			
12	IPSL-CM6A-					
	LR					





Table 4. Cross-scenario, KKZ-selected ESM subsets for Canada and each of its subregions, from the CanDCS-M6 downscaled CMIP6 scenarios. Models in italics differ between the CanDCS-U6 and CanDCS-M6 subsets.

	CANADA	Eastern	Ontario	Prairies	British	Northern
		Canada			Columbia	Canada
1	BCC-CSM2-	GFDL-ESM4	BCC-CSM2-	BCC-CSM2-	TaiESM1	CanESM5
	MR		MR	MR		
2	NorESM2-	NorESM2-	UKESM1-0-	EC-Earth3-	NorESM2-	INM-CM5-0
	LM	LM	LL	Veg	LM	
3	UKESM1-0-	MIROC-ES2L	NorESM2-	UKESM1-0-	CNRM-	NorESM2-
	LL		LM	LL	ESM2-1	LM
4	MRI-ESM2-0	FGOALS-g3	MIROC-ES2L	NorESM2-	MPI-ESM1-	MPI-ESM1-
				LM	2-HR	2-HR
5	MPI-ESM1-	MRI-ESM2-0	ACCESS-	FGOALS-g3	FGOALS-g3	MRI-ESM2-0
	2-HR		ESM1-5			
6	EC-Earth3-	UKESM1-0-	EC-Earth3-	INM-CM5-0	UKESM1-0-	ACCESS-
	Veg	LL	Veg		LL	ESM1-5
7	MIROC-ES2L	EC-Earth3-	INM-CM5-0	MRI-ESM2-0	MIROC-ES2L	TaiESM1
		Veg				
8	INM-CM5-0	INM-CM5-0	CanESM5	TaiESM1	MRI-ESM2-0	CMCC-ESM2
9	CMCC-ESM2	BCC-CSM2-	CNRM-	CNRM-	IPSL-CM6A-	BCC-CSM2-
		MR	ESM2-1	ESM2-1	LR	MR
10	FGOALS-g3	IPSL-CM6A-	MRI-ESM2-0	ACCESS-		EC-Earth3-
		LR		ESM1-5		Veg
11	TaiESM1		GFDL-ESM4			
12	IPSL-CM6A-					
	LR					

6. Summary

Representative subsets of the two downscaled sets of CMIP6 ESMs have been selected using the KKZ method and a novel cross-scenario selection procedure. The resulting national and regional subsets should simplify the use of CMIP6 downscaled results distributed through PCIC's Statistically Downscaled Climate Scenarios data portal.





References

Brunner, L., Pendergrass, A. G., Lehner, F., Merrifield, A. L., Lorenz, R., & Knutti, R., 2020: Reduced global warming from CMIP6 projections when weighting models by performance and independence. *Earth System Dyn.*, 11, 995-1012, <u>https://doi.org/10.5194/esd-11-995-2020</u>.

Cannon, A. J., 2015: Selecting ESM scenarios that span the range of changes in a multimodel ensemble: Application to CMIP5 climate extremes indices. *J. Climate*, **28**(3), 1260-1267, <u>https://doi.org/10.1175/JCLI-D-14-00636.1</u>.

Hausfather, Z., Marvel, K., Schmidt, G. A., Nielsen-Gammon, J. W., & Zelinka, M., 2022: Climate simulations: recognize the 'hot model'problem. *Nature*, 605(7908), 26-29, doi: 10.1038/d41586-022-01192-2.

IPCC, 2021: Annex II: Models [Gutiérrez, J M., A.-M. Tréguier (eds.)]. In *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* [Masson-Delmotte, V., P. Zhai, A. Pirani, S.L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M.I. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J.B.R. Matthews, T.K. Maycock, T. Waterfield, O. Yelekçi, R. Yu, and B. Zhou (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, pp. 2087–2138, doi:10.1017/9781009157896.016.

Katsavounidis, I., C.-C. J. Kuo, and Z. Zhang, 1994: A new initialization technique for generalized Lloyd iteration. *IEEE Signal Process. Lett.*, **1**, 144–146, doi:10.1109/97.329844. Pacific Climate Impacts Consortium, University of Victoria, 2019: Statistically Downscaled Climate Scenarios. Accessed from https://pacificclimate.org/data/statistically-downscaled-climate-scenarios.

Seneviratne, S. I., Donat, M. G., Pitman, A. J., Knutti, R., & Wilby, R. L., 2016: Allowable CO₂ emissions based on regional and impact-related climate targets. *Nature*, 529(7587), 477–483. https://doi.org/10.1038/ nature16542.

Sobie, S. R., Zwiers, F. W., & Curry, C. L., 2021: Climate Model Projections for Canada: A Comparison of CMIP5 and CMIP6. *Atmosphere-Ocean*, **59**(4-5), 269-284, <u>https://www.tandfonline.com/doi/full/10.1080/07055900.2021.2011103</u>.

Zhang, X., L. Alexander, G. C. Hegerl, P. Jones, A. K. Tank, T. C. Peterson, B. Trewin, and F. W. Zwiers, 2011: Indices for monitoring changes in extremes based on daily temperature and precipitation data. *Wiley Interdiscip. Rev.: Climate Change*, **2**, 851–870, doi:10.1002/wcc.147.