An efficient method for choosing the best sub-ensemble of climate models for ΔT Projections

Conor I. Anderson*[1], Ken Butler[2] and William A. Gough[1]

[1]Department of Physical and Environmental Sciences, UTSC
[2]Department of Computer and Mathematical Sciences, UTSC

March 11, 2018

Abstract

Climate scientists produce projections of future climate using Global Climate Models (GCMs), however, for a regional climate study, it is sometimes inconvenient or otherwise impractical to analyze each of the available models individually. Often, scientists will instead perform an 'ensemble' study, by averaging the output of all available models. Recent studies have shown, however, that representative sub-ensembles may outperform individual models and full ensembles alike. We propose an efficient method for the selection of representative sub-ensemble members for ΔT projections based on the sub-ensemble's ability to reproduce an observed climate baseline. We used our method to validate GCM reproduction of long-term baseline temperature averages at Toronto and Montreal, and found that sub-ensembles consistently outperformed individual models and total model ensembles when tested for individual variables and stations. Furthermore, sub-ensembles were more effective at reproducing the baseline average across combinations of variables and stations. Our method, derived and developed in the R programming language for statistical computing, provides a fast, computationally efficient solution to the selection of model members of selective, representative sub-ensembles.

Keywords: *general circulation model, sub-ensemble, model selection, ΔT, climate change*

1 Introduction

Climate scientists produce projections of future climate using Global Climate Models (GCMs). Individual climate models are developed by distinct institutions, from universities to government research agencies. Any given modelling group has its own unique series of priorities, interests, and foci, which may partly explain the diversity of climate models that are available. The international modelling community, however, has made strides to improve the interoperability of individual models. The fifth phase of the Combined Model Intercomparison Project, for instance, sought to

*conor.anderson@utoronto.ca

produce a freely available, state-of-the-art multimodel dataset (Taylor, Stouffer, & Meehl, 2011). The CMIP5 project involves more than 20 modelling groups and over 50 models (Taylor et al., 2011), providing scientists and impact assessors with a diverse range of possible futures for analysis.

In a regional applied climate study, it is sometimes inconvenient or otherwise impractical to report the projections of all available models. Using a wide range of models can increase the time and computing resources required to produce a study, and can produce results that are confusing to non-specialized audiences, an obstacle to the production of ‘policy-ready’, applied climate science. In these situations, it is often better to use select individual models, total model ensembles, or some sub-ensemble of all available models.

1.1 Approaches to Selecting Climate Models

Scientists may rely one of a number of methods to select which models to use for the evaluation of climate impacts. Fenech, Comer, and Gough (2007) describe three approaches: the use of extremes; model validation; and the use of multi-model ensembles. Analysis of maximum and minimum extremes implies using both the smallest and the largest projected changes as reference limits for policy planning. In this sense, the analysis is based on the extremes of the full range of possible futures projected by the available GCMs.

Individual models are often validated based on their ability to reproduce past and present climate (e.g. Ragettli, Pellicciotti, Bordoy, & Immerzeel, 2013; Xie, Gough, Tam, Zhao, & Wu, 2013; Hewer & Gough, 2016). In such a study, climatologists work under the assumption that a model’s ability to reproduce an observed climatic baseline will be representative of the model’s ability to project future climate. This is, at minimum, a necessary requirement, however, there are still challenges to this approach, such as that posed by equifinality; Gough (2001) showed that changes in the parameterization of model runs can greatly change the resulting future projections, even if the baseline is reproduced equally well by each run.

Multimodel ensembles—wherein the output from some selection of models is averaged, and that average is used to project future climate—often outperform individual models (Pierce, Barnett, Santer, & Gleckler, 2009). The ensemble approach somewhat addresses the varying strengths and weaknesses of individual models described above, and assumes that model biases will be cancelled within the averaged output (Knutti, Furrer, Tebaldi, Cermak, & Meehl, 2010). Indeed, the ensemble approach does tend to cancel offsetting errors present in individual models (Pierce et al., 2009), and account for differences in the parameterization of the models (Knutti et al., 2010). In this sense, some studies may choose to make use of a full ensemble of all available models, to account for the entire range of available projection data (e.g. Gough, Anderson, & Herod, 2016).

Often, a ‘sub-ensemble‘, or a subset of the available climate models both out-performs any individual model, and outperforms the full ensemble of models (Knutti et al., 2010). Lutz et al. (2016) provide numerous examples of the various methods that exist to select members of a multi-model ensemble. These include, among others: the KKZ algorithm developed by Katsavounidis, Jay Kuo, and Zhang (1994) and applied to climatological extremes by Cannon (2014); the detection of inter-model similarities based on principal component analysis and cluster analysis to shrink ensembles to statistically independent, representative simulations (Mendlik & Gobiet, 2016); and advanced, three-stage distillation of climate ensemble members based on projected changes to precipitation and temperature, projected changes in indices of climatic extremes, and each models ability to reproduce baseline climate (Lutz et al., 2016).

The above methods are statistically rigorous, and require a great deal of expert knowledge to implement. They are also based on information that is lost in the long-term averages computed for a ΔT-based study, namely, climatic extremes, seasonality and inter-annual variability. At the UTSC climate lab, Hewer and Gough (2016), opt for a simpler selection method. Their 'selective ensemble' approach used the three models that showed the smallest absolute difference from the observed baseline, providing a statistically-sound, yet methodologically simple solution to the problem of model selection. As such, their approach is accessible to a wide range of end users of GCM-data.

The ranking of models under the assumption that all models provide equally-plausible scenarios is both computationally efficient, and clear to end-users. It should be noted, however, that the top ranked models do not necessarily provide the multi-model ensemble that is closest to the observed baseline. When models are ranked by their absolute difference from the observed baseline, the sign of that difference (positive or negative) is lost, as such, the ensemble mean of the top three models may be further from zero than some combination of models with larger deviations that sum to zero.

It is critically important to select representative climate models, however, the myriad methods by which this process can be achieved often require advanced computation. It is in this context that we sought to develop an efficient, semi-automated, and novel method of selecting model members of a sub-ensemble for a ΔT study, based on that sub-ensemble's ability to reproduce an observed long-term baseline average. Our approach is implemented in the R language for statistical computing (R Core Team, 2017), a well-documented and freely-available programming language used by scientists and academics worldwide.

1.2 Selecting the best sub-ensemble

The most straight-forward and simple method of selecting the best sub-ensemble is to search for it by iteratively, testing all possible combinations of the available models, starting with single models ($k = 1$), and working upwards with larger ensembles up to a full ensemble such as $k = 2, k = 3, \ldots, k = n$, where k is the number of models in each sub-ensemble, and n is the number of available climate models. While theoretically simple, this method is computationally inefficient due to the combinatorial explosion of possible sub-sensemble configurations that that must be considered. There are 41 CMIP5 models in the 'r1i1p1' ensemble with baseline data for minimum temperature at surface over the 1981–2010 tridecade (details on how we obtained this data are in Section 2.1). For these 41 models, there are nearly 2.2 trillion distinct combinations that can be made. Even high-end modern systems are likely to have issues computing these combinations, let alone the necessary calculation of mean, standard deviation, and absolute difference from the observed values that will indicate which combination is the most apt at reproducing the observed climate. Given the technical limitations of the above method, it is not a practical solution. Indeed, a technically efficient algorithm is required to quickly provide an accurate and replicable estimate of the best sub-ensemble for a given station or set of stations.

Given that we consider the best sub-ensemble to be the group of models that most closely reproduces the baseline average measure of the variable of interest, such as minimum temperature, we can use integer programming to construct a sub-ensemble that minimizes this absolute difference. First, the model-generated projections of baseline temperature are numbered and denoted as x_i, where i goes from 1 to n (1 to 41 in the above example). The Boolean decision of whether a model will be included in a sub-ensemble, is represented mathematically as $a_i = 1$ or $a_i = 0$ according to whether model i is in the sub-ensemble or not.

We will let t denote the target variable, which in this case is the minimum temperature for our area of study over the baseline period, calculated as a 30-year mean temperature. For a set of n climate models, therefore, we will choose $a_i \in \{0,1\}$ so as to minimize:

$$\left| \frac{\sum_i x_i a_i}{\sum_i a_i} - t \right| \tag{1}$$

This equation, however, is difficult to solve because it is non-linear, a problem that is resolved by focusing on one particular sub-ensemble size at a time—that is, with $\sum_i a_i = k$ for fixed k. Equation 1 is then equivalent to Equation 2.

$$\text{Minimize } \left| \sum_i x_i a_i - kt \right| \text{ such that } \sum_i a_i = k \tag{2}$$

For a fixed k, this is linear with a single constraint (that $\sum_i a_i$ equal k), however, the absolute value still poses a challenge. We can make the equation more readily solvable by introducing a new variable and adding two additional constraints (Chiang, 2007). We add upper and lower constraints as s and $-s$, so that Equation 2 becomes Equation 3.

$$\begin{aligned} \text{Minimize} \quad & s \\ \text{such that } \sum_i x_i a_i - kt & \leq s \\ \sum_i x_i a_i - kt & \geq -s \\ \sum_i a_i & = k \end{aligned} \tag{3}$$

The above equation has exactly three constraints, and $n+1$ variables to minimize over, $a_1 \dots a_n$ and s. Finally, since k and t are constants and s is a variable, we rewrite Equation 3 as Equation 4.

$$\begin{aligned} \text{Minimize} \quad & s \\ \text{such that } \sum_i x_i a_i - s & \leq kt \\ \sum_i x_i a_i + s & \geq kt \\ \sum_i a_i & = k \end{aligned} \tag{4}$$

A key limitation of the selection of individual models or sub-ensembles for ΔT studies is that the best model for one variable is often different than the best model for another. At Toronto Pearson, for instance, the model ACCESS1.3 best reproduces the baseline average maximum temperature (T_{max}), with an absolute difference of 0.11 °C. The minimum temperature (T_{min}) is better-reproduced by the Australian CSIRO Mk. 3.6 model, with an absolute difference of 0.02 °C. We can modify Equation 4 to produce a sub-ensemble that best reproduces the baseline climate over one or more variables by adding additional constraints. For instance, if we want to find the sub-ensemble that best reproduces both T_{max} (we'll donote this as x) and T_{min} (we'll denote this as y), then we could use five constraints, as in Equation 5.

$$\begin{aligned} \text{Minimize} \quad & s \\ \text{such that} \sum_i x_i a_i - s & \leq k t_x \\ \sum_i x_i a_i + s & \geq k t_x \\ \sum_i y_i a_i - s & \leq k t_y \\ \sum_i y_i a_i + s & \geq k t_y \\ \sum_i a_i & = k \end{aligned} \tag{5}$$

The above version of the equation is exciting, because we can add as many constraints as we have variables and/or stations. In order to achieve this, we have created the 'subensemble' function in R (Anderson, 2017) that dynamically adds constraints based on the input.

2 Application to climate analysis in Toronto and Montreal

The power of this function can be explored further by means of a brief case study. The city of Toronto, Ontario, is the largest city in Canada. Toronto is characterized by a humid continental climate, which is influenced by the interaction of polar and tropical air masses due to the city's location near the latitudinal frontier of the northern polar vortex, which influences temperatures in all seasons (e.g. Anderson and Gough 2017, winter; Gough and Sokappadu 2015, summer). Toronto's downtown is characterized by the lake breeze from Lake Ontario, one of the Laurentian Great Lakes (Gough & Rozanov, 2001), but this effect is not present at the Pearson Airport station (Mohsin & Gough, 2012). Montreal, Quebec, is located to the northeast of Toronto in the same humid continental climate zone. The city is built on the Island of Montreal, at the confluence of the Ottawa and Saint Lawrence rivers. Montreal's minimum temperature is colder, on average, than Toronto's, but average maximums are similar for the two cities. Like Toronto, Montreal exhibits an urban heat island (Gough & Rozanov, 2001; Tam, Gough, & Mohsin, 2015). The city does not, however, experience a lake breeze (Gough & Rozanov, 2001). We will select three stations from each city to test Equation 5.

2.1 Data and methodology

We will use data provided by Environment and Climate Change Canada (formerly Environment Canada) for the period from 1981 to 2010 for six stations at Toronto and Montreal. At Toronto: Toronto City (previously "Toronto", station 5051, 1981–2002; station 31688, 2003–2010), is located on the University of Toronto's St. George campus and has the longest climate record in Canada. The Toronto Island station (5085, 1981 to January 2006; 30247, "Toronto City Centre", February 2006 to 2010) is located on Toronto Island at the Billy Bishop Toronto City Airport, and Toronto Pearson (5097) at the Lester B. Pearson International Airport in Mississauga, approximately 20 km west of the downtown core. At Montreal: the Pierre Elliot Trudeau International Airport (station 5415), the St. Hubert Airport (5490, 1981–2004; 30170, 2005–2009; 48374, 2010) and Ste. Genevieve (5484), located on the Island of Montreal near the Prairies river. We have selected a city station, local airport, and international airport in each of our two municipalities. The stations are shown in Figure 1.

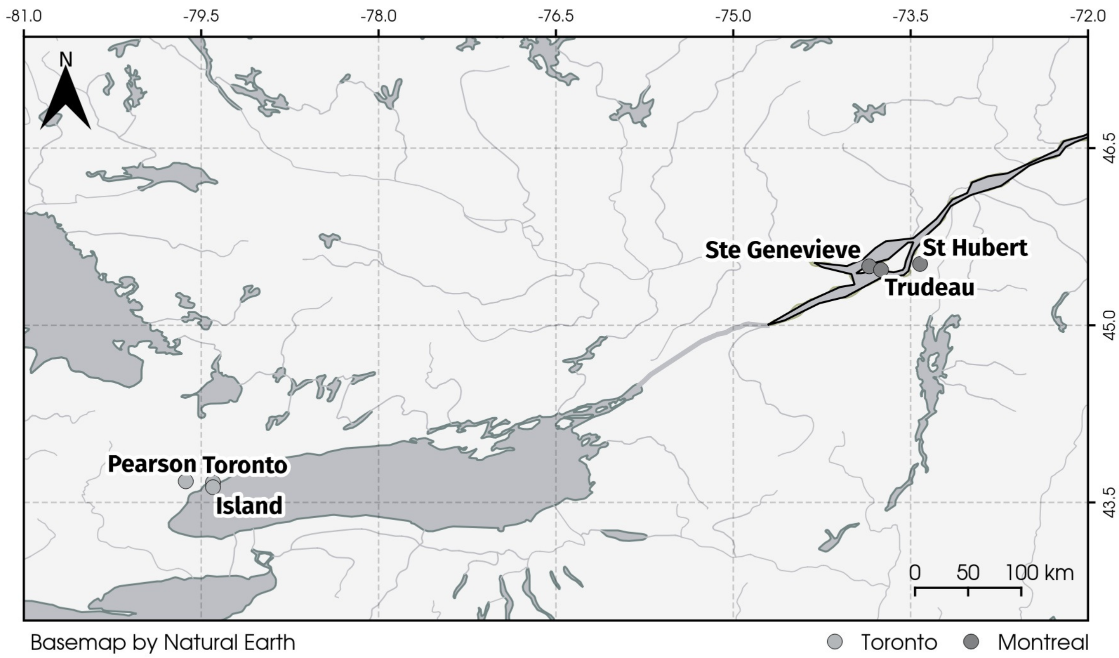

Figure 1: Map of study stations at Toronto and Montreal.

Daily data was acquired form Environment and Climate Change Canada for each of these stations from January 1981 to December 2010—30 years of data. The daily data was aggregated to monthly data, setting any month with more than 20% missing values as missing. These monthly data sets were used to calculate the mean 30-year baseline temperature for each station, omitting any year that was missing two months or more. These long-term baseline averages are presented in Table 1. Under this methodology, Ste. Genevieve was missing the most data, with all years from 1991 to 1998 set as missing. Trudeau and St. Hubert were missing one year each, 1993 and 1994, respectively. Toronto Island was missing annual data for 1994–95 and 2010.

Table 1: Thirty-year baseline averages for T_{max}, and T_{min} (°C) at Toronto and Montreal (1981–2010)

City	Station	Coordinates	T_{min}	$\sigma_{T\,min}$	T_{max}	$\sigma_{T\,max}$
	Toronto	43.67 °N, 79.4 °W	5.8	0.75	13.0	0.84
Toronto	Island	43.63 °N, 79.4 °W	5.0	0.71	12.2	0.79
	Pearson	43.68 °N, 79.63 °W	3.4	1.07	13.0	0.90
	Ste Genevieve	45.5 °N, 73.85 °W	2.6	0.82	10.9	0.73
Montreal	St Hubert	45.52 °N, 73.42 °W	1.1	0.93	11.5	0.76
	Trudeau	45.47 °N, 73.75 °W	2.0	0.96	11.5	0.78

We acquired GCM output for 41 models in the 'r1i1p1' ensemble from The Earth System Grid Federation (Cinquini et al., 2014). These model outputs were processed on the '*Conjuntool*' platform (Anderson, 2018), a freely-available, open-source platform that we developed during the course of the present study, and that, at the time of writing, was still undergoing heavy development. Using the '*Conjuntool*', we obtained baseline hindcasts for the gridcell containing each of our six stations for the period from 1971 to 2000. These hindcasts were processed for each of the 41 models, and were returned as 30-year climatological averages.

The projected baseline average temperature, for each model is presented in Table 2. Most of the "historical" CMIP5 model output ends in December, 2005, with climate forcing applied to all dates from January 2006, onward. As such, some of our baseline values include the effects of one, or some combination of the four RCP scenarios. We have averaged the output for each model, to minimize the influence of any one scenario. The methodology we have derived earlier in this paper will help to choose which models to include in a representative sub-ensemble.

Our function calculates the absolute difference between the hindcasted baseline values and the observed baseline average temperature, ranking the differences by the Gough-Fenech confidence index (GFCI) to test the confidence of each model's projections (Fenech, 2009; Hewer & Gough, 2016). The GFCI is described by Equation 6.

$$GF = \frac{|T_{obs} - T_{se}|}{\sigma_{obs}} \tag{6}$$

Where T_{obs} is the observed baseline temperature; T_{se} is the average baseline projected by the sub-ensemble of models, and σ_{obs} is the standard deviation of the aggregate annual observed data. The GFCI, therefore, is the standardized absolute difference between the actual temperature and projected temperature. We will consider GF values below 1.0 to indicate moderate confidence; values below 0.5 will represent strong confidence, and values approaching 0 will represent the strongest confidence. Any value that is above 1.0, that is, more than one standard deviation away from the true baseline, will be considered to lack confidence.

2.2 Results

By way of example, the formatted output of the 'subensemble' function to find the best sub-ensemble for minimum temperature at Toronto City is presented in Table 3.

The individual model that best reproduced the baseline temperature in Toronto was model number 2—ACCESS1.3 from Australia's Commonwealth Scientific and Industrial Research Organisation and Bureau of Meteorology. The baseline temperature projected by ACCESS1.3 resulted in a residual of approximately -0.022 °C, while the full 41-model ensemble, on the other hand, had a residual of -2.4 °C.

Every multi-model sub-ensemble outperformed the full ensemble; sub-ensembles of between 2 and 11 models also outperformed the single best model. Those sub-ensembles all showed absolute differences from the observed mean that were less than the limits of the precision of the observed data, which is measured to one tenth of a degree Celcius. In this sense, any of these sub-ensembles could be selected as the theoretical best sub-ensemble. Likewise, all sub-ensembles of between 1 and 15 models passed the GFCI at $GF < 0.5$; sub-ensembles of 16 to 20 models showed moderate confidence. The sub-ensemble that was the closest, in absolute value, to the observed baseline was the 8-model sub-ensemble, which had a residual of -1.8×10^{-6} °C and a GF value of 2.5×10^{-6}.

Table 2: Projected historical baseline (1981–2010) temperature (°C) at Toronto and Montreal

No.	Model	T_{min} Toronto	Island	Pearson	Ste Genev.	St Hubert	Trudeau	T_{max} Toronto	Island	Pearson	Ste Genev.	St Hubert	Trudeau
1	ACCESS1-0[2,4]	4.7	4.7	4.7	2.5	2.5	2.5	11.8	11.8	11.8	9.9	9.9	9.9
2	ACCESS1-3[2,4]	5.8	5.8	5.8	3.8	3.8	3.8	13.1	13.1	13.1	10.9	10.9	10.9
3	bcc-csm1-1	2.2	2.2	2.2	-1.3	-1.3	-1.3	11.3	11.3	11.3	8.1	8.1	8.1
4	bcc-csm1-1-m	3.7	3.7	3.7	**0.7**	**0.6**	**0.7**	13.8	13.8	13.8	**11.3**	**11.0**	**11.3**
5	BNU-ESM[1,2,4]	0.4	0.4	0.4	-2.2	-2.2	-2.2	11.5	11.5	11.5	8.3	8.3	8.3
6	CanESM2[1,2,4]	7.1	7.1	7.1	3.1	3.1	3.1	17.4	17.4	17.4	13.0	13.0	13.0
7	CCSM4[1,2,3,4]	3.8	3.8	3.8	1.5	1.5	1.5	13.5	13.5	13.5	12.3	12.3	12.3
8	CESM1-BGC[2,4]	3.8	3.8	3.8	1.5	1.5	1.5	13.6	13.6	13.6	12.3	12.3	12.3
9	CESM1-CAM5[1,2,3,4]	2.3	2.3	2.3	-0.3	-0.3	-0.3	12.1	12.1	12.1	10.5	10.5	10.5
10	CMCC-CESM	1.6	1.6	1.6	-2.0	-2.0	-2.0	10.3	10.3	10.3	6.8	6.8	6.8
11	CMCC-CM[2,4]	3.1	3.1	3.1	0.8	0.8	0.8	11.1	11.1	11.1	9.4	9.4	9.4
12	CMCC-CMS[2,4]	3.1	3.1	3.1	-0.5	-0.5	-0.5	10.3	10.3	10.3	7.2	7.2	7.2
13	CNRM-CM5[1,2,4]	**1.8**	**1.8**	**1.7**	**1.0**	**0.4**	**0.4**	**11.7**	**11.7**	**11.4**	**10.5**	**10.1**	**10.1**
14	CSIRO-Mk3-6-0[1,2,3,4]	3.4	3.4	3.4	-0.7	-0.7	-0.7	10.7	10.7	10.7	9.5	9.5	9.5
15	EC-EARTH[2,4]	4.3	4.3	4.3	**1.3**	**1.1**	**1.3**	11.0	11.0	11.0	**9.9**	**9.8**	**9.9**
16	FGOALS-g2	-1.1	-1.1	-1.1	-3.3	-3.3	-3.3	8.3	8.3	8.3	6.1	6.1	6.1
17	FIO-ESM[1,2,3,4]	4.0	4.0	4.0	1.3	1.3	1.3	12.5	12.5	12.5	10.7	10.7	10.7
18	GFDL-CM3[1,2,3,4]	4.6	4.6	4.6	1.3	1.3	1.3	10.1	10.1	10.1	8.4	8.4	8.4
19	GFDL-ESM2G[1,2,3,4]	4.4	4.4	4.4	0.0	0.0	0.0	8.3	8.3	8.3	6.1	6.1	6.1
20	GFDL-ESM2M[1,2,3,4]	5.4	5.4	5.4	1.4	1.4	1.4	9.2	9.2	9.2	7.3	7.3	7.3
21	GISS-E2-H[1,2,3,4]	5.0	5.0	5.0	2.7	2.7	2.7	11.5	11.5	11.5	10.1	10.1	10.1
22	GISS-E2-H-CC	4.8	4.8	4.8	2.3	2.3	2.3	11.4	11.4	11.4	9.8	9.8	9.8
23	GISS-E2-R[1,2,3,4]	4.8	4.8	4.8	2.5	2.5	2.5	11.4	11.4	11.4	10.1	10.1	10.1
24	GISS-E2-R-CC	4.8	4.8	4.8	2.5	2.5	2.5	11.5	11.5	11.5	10.0	10.0	10.0
25	HadCM3	1.8	1.8	1.8	-0.2	-0.2	-0.2	10.8	10.8	10.8	9.0	9.0	9.0
26	HadGEM2-AO[1,2,3,4]	3.5	3.5	3.5	1.6	1.6	1.6	11.9	11.9	11.9	9.8	9.8	9.8
27	HadGEM2-CC[2,4]	0.9	0.9	0.9	-0.5	-0.5	-0.5	9.7	9.7	9.7	8.2	8.2	8.2
28	HadGEM2-ES[1,2,3,4]	2.5	2.5	2.5	0.8	0.8	0.8	11.2	11.2	11.2	9.5	9.5	9.5
29	inmcm4[2,4]	2.4	2.4	2.4	0.8	0.8	0.8	12.4	12.4	12.4	10.8	10.8	10.8
30	IPSL-CM5A-LR[1,2,3,4]	-0.8	-0.8	-0.8	**-8.6**	**-8.6**	**-6.1**	15.7	15.7	15.7	**15.9**	**15.9**	**17.8**
31	IPSL-CM5A-MR[1,2,3,4]	0.8	0.8	0.8	**-5.9**	**-6.1**	**-6.1**	19.9	19.9	19.9	**19.4**	**19.1**	**19.1**
32	IPSL-CM5B-LR[2,4]	-1.7	-1.7	-1.7	**-12.2**	**-12.2**	**-9.3**	15.0	15.0	15.0	**14.8**	**14.8**	**16.7**
33	MIROC-ESM[1,2,3,4]	6.4	6.4	6.4	3.8	3.8	3.8	14.1	14.1	14.1	11.5	11.5	11.5
34	MIROC-ESM-CHEM[1,2,3,4]	5.9	5.9	5.9	3.5	3.5	3.5	13.8	13.8	13.8	11.2	11.2	11.2
35	MIROC4h	**8.3**	**8.3**	**7.1**	**3.9**	**3.9**	**4.6**	**14.6**	**14.6**	**15.0**	**11.9**	**11.9**	**12.6**
36	MIROC5	**6.8**	**6.8**	**7.2**	**2.7**	**2.2**	**2.2**	**13.5**	**13.5**	**13.7**	10.6	10.6	10.6
37	MPI-ESM-LR[2,4]	3.8	3.8	3.8	0.8	0.8	0.8	11.0	11.0	11.0	8.4	8.4	8.4
38	MPI-ESM-MR[2,4]	3.9	3.9	3.9	1.1	1.1	1.1	11.4	11.4	11.4	9.0	9.0	9.0
39	MRI-CGCM3[1,2,3,4]	3.9	3.9	3.9	**0.2**	**0.1**	**0.2**	10.3	10.3	10.3	**9.0**	**8.9**	**9.0**
40	MRI-ESM1	4.1	4.1	4.1	**0.4**	**0.3**	**0.4**	10.5	10.5	10.5	9.1	9.1	9.1
41	NorESM1-M[1,2,3,4]	1.4	1.4	1.4	**-2.1**	**-1.7**	**0.5**	11.1	11.1	11.1	**9.6**	**9.3**	**11.8**
	<ENSEMBLE>	**3.5**	**3.5**	**3.4**	**0.2**	**0.2**	**0.4**	12.1	12.1	12.1	**10.2**	**10.1**	**10.3**

The above table presents the projected baseline (1981–2010) average temperature for 41 models at 6 stations. In most cases, the stations at Toronto or at Montreal fall within the same grid cell, so the projected baseline value is similar between stations. Cases where different values are projected are due to varying resolutions between models; these values have been marked in **bold**. Most models project climate change impacts from January 2006 onward, hence some of these baseline values are partially forced with (1) RCP2.6, (2) RCP4.5, (3) RCP6.0, and/or (4) RCP8.5; these cases are denoted in superscript. For cases where more than one RCP scenario was used, the value above denotes the average.

Table 3: Output of `subensemble` function for 41 model runs of average minimum temperature at Toronto (1981–2010)

k	Res. (°C)	GF	a1	a2	a3	a4	a5	a6	a7	a8	a9	a10	a11	a12	a13	a14	a15	a16	a17	a18	a19	a20	a21	a22	a23	a24	a25	a26	a27	a28	a29	a30	a31	a32	a33	a34	a35	a36	a37	a38	a39	a40	a41
1	-0.022	0.030	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	-0.0055	0.0074	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0.0014	0.0018	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	1	0	0	0	0	0
4	-0.00034	0.00046	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0
5	-4.6×10^{-5}	6.1×10^{-5}	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0
6	-3.7×10^{-6}	5.0×10^{-6}	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	0
7	-8.5×10^{-6}	1.1×10^{-5}	1	1	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
8	-1.8×10^{-6}	2.5×10^{-6}	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	1	1	1	0	0	0	0
9	-1.5×10^{-5}	2.0×10^{-5}	1	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	1	0	0	0
10	-1.2×10^{-5}	1.6×10^{-5}	1	1	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0	0
11	-1.3×10^{-5}	1.7×10^{-5}	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1	1	1	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0	0
12	-0.030	0.039	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0	0
13	-0.12	0.16	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	1	1	1	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0	0
14	-0.22	0.29	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0	0
15	-0.31	0.41	1	1	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0	0
16	-0.40	0.53	1	1	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0	1	0
17	-0.48	0.64	1	1	0	0	0	1	0	0	0	0	0	0	0	0	1	0	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0	1	0
18	-0.56	0.75	1	1	0	0	0	1	0	0	0	0	0	0	0	0	1	0	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	1	1	1	1	0	1	0	1	0
19	-0.64	0.85	1	1	0	0	0	1	0	0	0	0	0	0	0	0	1	0	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	1	1	1	1	0	1	1	1	0
20	-0.71	0.94	1	1	0	0	0	1	0	0	0	0	0	0	0	0	1	0	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	0
21	-0.77	1.0	1	1	0	0	0	1	0	1	0	0	0	0	0	0	1	0	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	0
22	-0.83	1.1	1	1	0	0	0	1	1	1	0	0	0	0	0	0	1	0	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	0
23	-0.89	1.2	1	1	0	1	0	1	1	1	0	0	0	0	0	0	1	0	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	0
24	-1.0	1.3	1	1	0	1	0	1	1	1	0	0	0	0	0	0	1	0	1	1	1	1	1	1	1	1	0	1	0	0	0	0	0	0	1	1	1	1	1	1	1	1	0
25	-1.0	1.3	1	1	0	1	0	1	1	1	0	0	0	0	0	1	1	0	1	1	1	1	1	1	1	1	0	1	0	0	0	0	0	0	1	1	1	1	1	1	1	1	0
26	-1.1	1.4	1	1	0	1	0	1	1	1	0	0	1	0	0	1	1	0	1	1	1	1	1	1	1	1	0	1	0	0	0	0	0	0	1	1	1	1	1	1	1	1	0
27	-1.1	1.5	1	1	0	1	0	1	1	1	0	0	1	1	0	1	1	0	1	1	1	1	1	1	1	1	0	1	0	0	0	0	0	0	1	1	1	1	1	1	1	1	0
28	-1.2	1.6	1	1	0	1	0	1	1	1	0	0	1	1	0	1	1	0	1	1	1	1	1	1	1	1	0	1	0	1	0	0	0	0	1	1	1	1	1	1	1	1	0
29	-1.3	1.7	1	1	0	1	0	1	1	1	0	0	1	1	0	1	1	0	1	1	1	1	1	1	1	1	0	1	0	1	1	0	0	0	1	1	1	1	1	1	1	1	0
30	-1.4	1.8	1	1	0	1	0	1	1	1	1	0	1	1	0	1	1	0	1	1	1	1	1	1	1	1	0	1	0	1	1	0	0	0	1	1	1	1	1	1	1	1	0
31	-1.4	1.9	1	1	1	1	0	1	1	1	1	0	1	1	0	1	1	0	1	1	1	1	1	1	1	1	0	1	0	1	1	0	0	0	1	1	1	1	1	1	1	1	0
32	-1.5	2.0	1	1	1	1	0	1	1	1	1	0	1	1	0	1	1	0	1	1	1	1	1	1	1	1	1	1	0	1	1	0	0	0	1	1	1	1	1	1	1	1	0
33	-1.6	2.1	1	1	1	1	0	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	1	1	0	0	0	1	1	1	1	1	1	1	1	0
34	-1.7	2.2	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	1	1	0	0	0	1	1	1	1	1	1	1	1	0
35	-1.8	2.4	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	1	1	0	0	0	1	1	1	1	1	1	1	1	1
36	-1.8	2.5	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	1	1	1	1	1	1	1	1	1
37	-1.9	2.6	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	1	1	1	1	1	1	1	1	1
38	-2.0	2.7	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	1	1	1	1	1	1	1	1	1
39	-2.1	2.9	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1
40	-2.3	3.0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1
41	-2.4	3.2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

We can perform the same basic analysis for a two-variable test, using both maximum and minimum temperatures at Toronto. Select results of this test are presented in Table 4. Sub-ensembles out-performed both the top individual model (the Australian ACCESS1.3, average GF of 0.066) and the full 41-model ensemble (Average GF of 2.2). Sub-ensembles of between 1 and 15 models again merited high confidence, with GF values below 0.5; sub-ensembles between 16 and 21 models merited moderate confidence. The top three sub-ensembles all showed absolute differences below the precision of the observed data, and minute GFCI values. The best sub-ensemble used six models, with residuals for T_{max} and T_{min} of 0.0013 °C and −0.0021 °C, respectively, and an average GF value of 0.0021.

Table 4: Best sub-ensembles for maximum and minimum temperature at Toronto (1981–2010)

k	Res. T_{max} (°C)	Res. T_{min} (°C)	GF T_{max}	GF T_{min}	Avg. GF	Models Used
6	0.0013	-0.0021	0.0015	0.0028	0.0021	6,15,18,22,34,35
8	-0.0021	-0.0012	0.0025	0.0016	0.0021	6,8,20,22,33,35,36,37
9	0.0018	-0.0027	0.0021	0.0035	0.0028	4,6,15,18,20,33,34,35,36
1	0.086	-0.022	0.10	0.03	0.066	2 (ACCESS1.3)
41	-0.99	-2.4	1.2	3.2	2.2	ALL

As we add more variables and stations to our request, the reliability of our sub-ensembles decreases substantially. When we use both T_{min} and T_{max} at all three Toronto stations, the average GFCI value never drops below the 0.5 threshold of high confidence. For this six-constraint experiment, sub-ensembles, again, outperformed both the top-performing single model (ACCESS1.0, Average GF of 1.1) and the full ensemble (Average GF of 1.4). Our top-performing sub-ensemble was composed of just two models, CMCC-CESM and MIROC4h, however, with an average GF score of 0.59, this sub-ensemble failed to garner high confidence under the conditions of the GFCI. Only the scores at Toronto Island garnered confidence with GFCI values below 0.5. The score for minimum temperature at Toronto City did not meet our threshold for moderate confidence (GF <1.0) for any level of k. The individual scores for T_{max} and T_{min} at each station are summarized in Table 5.

Table 5: Statistical summary for the optimal 2-model sub-ensemble for T_{min} and T_{max} across three Toronto stations

Station	Res. T_{min} (°C)	Res. T_{max} (°C)	GF. T_{min}	GF. T_{max}	Avg. GF
Toronto	-0.91	-0.61	1.2	0.73	
Pearson	0.98	-0.38	0.92	0.42	0.59
Island	-0.017	0.20	0.024	0.25	

Our results were more encouraging for Montreal. When testing for both T_{min} and T_{max} across our three Montreal stations, we found 6 sub-ensembles that, on average, showed GF values below 0.5. In addition, all sub-ensemble sizes from 2 to 22 reproduced each of the station–variable baseline averages with GF values below 1.0, and an average GF value below 1.0. Once, again, we found that the total-model ensemble did the worst job of reproducing long-term baseline averages across the six station–variable pairs (Avg. GF of 1.7). Sub-ensembles also out-performed the single best model, (MIROC5, Avg. GF of 0.7). Our best sub-ensemble (Avg. GF of 0.42) was the one with 2 models (CESM1-BGC, and MIROC5). The results for this sub-ensemble are summarized in Table 6.

Table 6: Statistical summary for the optimal 2-model sub-ensemble for T_{min} and T_{max} across three Montreal stations

Station	Res. T_{min} (°C)	Res. T_{max} (°C)	GF. T_{min}	GF. T_{max}	Avg. GF
Ste. Genevieve	-0.68	0.53	0.81	0.80	
St Hubert	0.68	-0.24	0.80	0.31	0.46
Trudeau	0.034	-0.0069	0.036	0.0084	

Finally, we performed a 12-constraint test, searching for the best sub-ensemble for both T_{min} and T_{max}, across all six sites in the two municipalities. The single best model was GISS-E2-R-CC, however this model failed to reproduce baseline values to a GF value of below 1.0 for 9 of our 12 cases, and had an average GF value of 1.2. The 41-model total ensemble performed the poorest, with an average GF value of 1.5. Sub-ensembles out-performed the single best model and the total model ensemble, a consistent finding across all of our tests. In this case, the best sub-ensemble contained just two models, CNRM-CM5 and MIROC4h, showing an average GF value of 0.5. This sub-ensemble reproduced six of our variables with high confidence, and three with moderate confidence, but failed to reproduce baseline maximum temperature at Toronto Island or baseline minimum temperature at Toronto City or St. Hubert. The results for the individual stations and variables are shown in Table 7.

Table 7: Statistical summary for the optimal 2-model sub-ensemble for T_{min} and T_{max} across three Toronto and three Montreal stations

City	Station	Res. T_{min} (°C)	Res. T_{max} (°C)	GF. T_{min}	GF. T_{max}	Avg. GF
	Toronto	-0.77	0.10	1.0	0.12	
Toronto	Pearson	1.1	0.19	0.99	0.21	
	Island	0.12	0.91	0.16	1.2	0.57
	Ste Genevieve	0.36	0.36	0.20	0.49	
Montreal	St Hubert	1.0	-0.50	1.1	0.66	
	Trudeau	0.49	-0.15	0.50	0.19	

3 Discussion

Our method is unbiased, minimizing the absolute difference between the baseline observed temperature and the hindcast temperature provided by each sub-ensemble, In our tests, however, our algorithm produced sub-ensembles with higher confidence for Montreal, than for Toronto. The relatively coarse scale of the GCMs in our test set can partially explain this performance difference. Toronto is characterized by pronounced micro-climates related to the lake effect and the urban heat island (Gough & Rozanov, 2001). The urban heat island (UHI), which is most notable for minimum temperatures, is strongest at Toronto City, but is also present at Pearson (Mohsin & Gough, 2012). Toronto Island is unique among our three Toronto stations in that it is less developed, and, therefore, less impacted by the UHI, and is surrounded by water. When we projected for T_{max} and T_{min} across the three stations, our sub-ensemble favoured higher performance for T_{min} at Toronto Island. Notwithstanding our best model, larger sub-ensembles (especially with k between 16 and 41) favoured T_{max} at Toronto Island. In the case of Toronto, it seems that Toronto Island contributed disproportionately to the minimization of s. The presence of the Great Lakes also plays a role in the accuracy of projections over Toronto. Of 54 CMIP5 models assessed by the Great Lakes Ensemble Project, just twelve had a fraction of a grid cell that was not defined as land, and only eight provide sea surface temperature for at least part of one of the five Laurentian Great Lakes (Great Lakes Integrated Sciences and Assessments program (GLISA), 2017).

Among our sites in Montreal, Ste. Genevieve and St. Hubert do not experience a lake breeze. Trudeau is located near Lac St. Louis, a widening of the St. Lawrence; this body of water could produce a muted lake breeze, but we have not found literature dedicated to the subject. More notably, in contrast to our heavily urbanized sites at Toronto, our Montreal stations are less impacted by the urban heat island affect. Both Trudeau and Ste. Genevieve are located in the largely residential west end of Montreal. St. Hubert is to the east of the island, in the suburban city of Longueuil. Our optimal two-model sub-ensemble for Montreal favoured performance for T_{max} at Trudeau, however, this was less consistent than at Toronto, with differing sub-ensembles favouring Trudeau T_{min}, T_{max} at Ste. Genevieve, and T_{min} at St. Hubert. In the case of Montreal, the algorithm was driven showed no favouritism for a given site. The same observation held when we applied our analysis to both T_{max} and T_{min} across the six stations. The variability of the stations led to our algorithm to favour optimization, with no particular station appearing to disproportionately contribute to minimizing s.

We should highlight that in the present study, we used the 'r1i1p1' ensemble of CMIP5 results. Some models, such as the CCSM4 provide additional parameterization to account for the urban heat island (Oleson, 2011). There has also been an effort to provide high resolution down-scaled GCM results for urban areas (Lauwaet et al., 2015). As such, future research could apply our algorithm to an urban-optimized ensemble or models or to downscaled GCM output and compare the results with this paper.

4 Conclusions

We described and demonstrated our novel, efficient method for selecting a sub-ensemble of climate models based on the sub-ensemble's ability to reproduce long-term averages of an observed climatological baseline. We found that, in all cases, sub-ensembles out-perform both individual models and total-model ensembles. We were able to identify sub-ensembles that provided high-confidence in their reproduction of the baseline T_{max} and T_{min} at Toronto City. We were able to produce a number of sub-ensembles that reproduced the baseline for these variables across the three Toronto stations with moderate confidence. It is important to note, however, that none of these sub-ensembles reproduced baseline minimum temperature at Toronto City within our threshold of high confidence. At Montreal, a number of sub-ensemble configurations could reproduce the baseline T_{max} and T_{min} across the three stations with moderate to high confidence. We found a number of sub-ensemble sizes that could, on average, reproduce baselines to moderate confidence across all six stations and the two variables, but failed to identify any sub-ensemble size that reproduced each and every baseline station–variable pair with confidence. However, we highlight that in all tests, sub-ensembles exceeded the ability of single models and the total-model ensemble to reproduce baseline values.

Our method is based on minimizing the total difference between sub-ensemble averages and the observed baseline values. A future version of the method we used in the present study could therefore seek to introduce additional conditions to seek the sub-ensemble that first meets the GFCI limits for *all* station–variable pairs (if it exists), as opposed to minimizing the overall average difference.

Our method, along with the *Conjuntool*, is part of our efforts to provide, free, and open methodologies for climate analysis. Our function is easily used by non-technical end-users, however, depending on the expertise of the researcher, we suggest that this method form one component of more detailed methodologies for the generation of climate model sub-ensembles. Where full multi-variate, projected climate data is available, we suggest that our method be used in tandem with other methods such as performance analysis (e.g. Evans, Ji, Abramowitz, & Ekström, 2013) or cluster analysis (e.g. Cannon, 2014; Mendlik & Gobiet, 2016), such that the models that are available to our algorithm have a stronger degree of independence while still reproducing the range of climate data available in the full ensemble.

The mathematical optimization theory behind our method is not limited to climate science, or even to the natural sciences for that matter, we invite the research community to adopt, adapt, and improve our method for this, and other applications.

Acknowledgements

Thank you to Dr. Micah Hewer for valuable feedback on a draft of this manuscript.

References

Anderson, C. I. (2017). *claut: Functions from the University of Toronto Climate Lab*. R package version 0.1.2. Retrieved from https://gitlab.com/ConorIA/claut

Anderson, C. I. (2018). Conjuntool. Retrieved from https://gitlab.com/ConorIA/conjuntool

Anderson, C. I. & Gough, W. A. (2017). Evolution of Winter Temperature in Toronto, Ontario, Canada: A Case Study of Winters 2013/14 and 2014/15. *Journal of Climate*, *30*(14), 5361–5376. doi:10.1175/JCLI-D-16-0562.1

Cannon, A. J. (2014). 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. doi:10.1175/JCLI-D-14-00636.1

Chiang, M. (2007). Lecture 3A: Linear Programming. Retrieved from https://www.princeton.edu/~chiangm/ele539l3a.pdf

Cinquini, L., Crichton, D., Mattmann, C., Harney, J., Shipman, G., Wang, F., … Schweitzer, R. (2014). The Earth System Grid Federation: An open infrastructure for access to distributed geospatial data. *Future Generation Computer Systems*. Special Section: Intelligent Big Data Processing, *36*(Supplement C), 400–417. doi:10.1016/j.future.2013.07.002

Evans, J. P., Ji, F., Abramowitz, G., & Ekström, M. (2013). Optimally choosing small ensemble members to produce robust climate simulations. *Environmental Research Letters*, *8*(4), 044050. doi:10.1088/1748-9326/8/4/044050

Fenech, A. (2009). *Rapid assessment of the impacts of climate change: An integrated approach to understanding climate change in the Halton region of Ontario, Canada* (Unpublished Doctoral Dissertation, University of Toronto, Toronto, Ontario).

Fenech, A., Comer, N., & Gough, W. A. (2007). Selecting a Global Climate Model for Understanding Future Projections of Climate Change. In *Linking climate models to policy and decision-making* (pp. 133–145). Toronto: Environment Canada. Retrieved from http://projects.upei.ca/climate/files/2012/10/Book-1_Paper-5.pdf

Gough, W. A. (2001). Model tuning and its impact on modelled climate change response: Hudson Bay sea ice, a case study. *The Canadian Geographer / Le Géographe canadien*, *45*(2), 300–305. doi:10.1111/j.1541-0064.2001.tb01490.x

Gough, W. A., Anderson, V., & Herod, K. (2016). *Ontario Climate Change and Health Modelling Study: Report*. Ministry of Health and Long-Term Care Public Health Policy and Programs Branch. Ontario. doi:10.13140/RG.2.2.35542.96327

Gough, W. A. & Rozanov, Y. (2001). Aspects of Toronto's climate: Heat island and lake breeze. *Canadian Meteorological and Oceanographic Society Bulletin*, *29*(3), 67–71. Retrieved from http://cmosarchives.ca/Bulletin/b2903.pdf

Gough, W. A. & Sokappadu, S. (2015). Climate context of the cold summer of 2014 in Toronto, ON, Canada. *Theoretical and Applied Climatology*, 1–7. doi:10.1007/s00704-015-1571-2

Great Lakes Integrated Sciences and Assessments program (GLISA). (2017). *Great Lakes Ensemble March 2016 - March 2017 Progress Report*. Retrieved from http://www.glisa.umich.edu/media/files/Ensemble_Progress_Report_March_2017.pdf

Hewer, M. J. & Gough, W. A. (2016). Assessing the Impact of Projected Climate Change on Zoo Visitation in Toronto (Canada). *Journal of Geography and Geology*, *8*(2), 30–48. doi:10.5539/jgg.v8n2p30

Katsavounidis, I., Jay Kuo, C.-C., & Zhang, Z. (1994). A new initialization technique for generalized Lloyd iteration. *IEEE Signal Processing Letters*, *1*(10), 144–146. doi:10.1109/97.329844

Knutti, R., Furrer, R., Tebaldi, C., Cermak, J., & Meehl, G. A. (2010). Challenges in Combining Projections from Multiple Climate Models. *Journal of Climate*, *23*(10), 2739–2758. doi:10.1175/2009JCLI3361.1

Lauwaet, D., Hooyberghs, H., Maiheu, B., Lefebvre, W., Driesen, G., Van Looy, S., & De Ridder, K. (2015). Detailed Urban Heat Island Projections for Cities Worldwide: Dynamical Downscaling CMIP5 Global Climate Models. *Climate*, *3*(2), 391–415. doi:10.3390/cli3020391

Lutz, A. F., ter Maat, H. W., Biemans, H., Shrestha, A. B., Wester, P., & Immerzeel, W. W. (2016). Selecting representative climate models for climate change impact studies: An advanced envelope-based selection approach. *International Journal of Climatology*, *36*(12), 3988–4005. doi:10.1002/joc.4608

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. doi:10.1007/s10584-015-1582-0

Mohsin, T. & Gough, W. A. (2012). Characterization and estimation of urban heat island at Toronto: Impact of the choice of rural sites. *Theoretical and Applied Climatology*, *108*, 105–117. doi:10.1007/s00704-011-0516-7

Oleson, K. (2011). Contrasts between Urban and Rural Climate in CCSM4 CMIP5 Climate Change Scenarios. *Journal of Climate*, *25*(5), 1390–1412. doi:10.1175/JCLI-D-11-00098.1

Pierce, D. W., Barnett, T. P., Santer, B. D., & Gleckler, P. J. (2009). Selecting global climate models for regional climate change studies. *Proceedings of the National Academy of Sciences*, *106*(21), 8441–8446. doi:10.1073/pnas.0900094106

R Core Team. (2017). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Ragettli, S., Pellicciotti, F., Bordoy, R., & Immerzeel, W. W. (2013). Sources of uncertainty in modeling the glaciohydrological response of a Karakoram watershed to climate change. *Water Resources Research*, *49*(9), 6048–6066. doi:10.1002/wrcr.20450

Tam, B. Y., Gough, W. A., & Mohsin, T. (2015). The impact of urbanization and the urban heat island effect on day to day temperature variation. *Urban Climate*, *12*, 1–10. doi:10.1016/j.uclim.2014.12.004

Taylor, K. E., Stouffer, R. J., & Meehl, G. A. (2011). An Overview of CMIP5 and the Experiment Design. *Bulletin of the American Meteorological Society*, *93*(4), 485–498. doi:10.1175/BAMS-D-11-00094.1

Xie, C., Gough, W. A., Tam, A., Zhao, L., & Wu, T. (2013). Characteristics and Persistence of Relict High-Altitude Permafrost on Mahan Mountain, Loess Plateau, China. *Permafrost and Periglacial Processes*, *24*(3), 200–209. doi:10.1002/ppp.1776