



Performance of multi-model ensembles for the simulation of temperature variability over Ontario, Canada

Aly Al Samouly¹ · Chanh Nien Luong¹ · Zhong Li¹  · Spencer Smith² · Brian Baetz¹ · Maysara Ghaith¹

Received: 31 December 2017 / Accepted: 2 July 2018 / Published online: 9 July 2018
© Springer-Verlag GmbH Germany, part of Springer Nature 2018

Abstract

Climate ensembles utilize outputs from multiple climate models to estimate future climate patterns. These multi-model ensembles generally outperform individual climate models. In this paper, the performance of seven global climate model and regional climate model combinations were evaluated for Ontario, Canada. Two multi-model ensembles were developed and tested, one based on the mean of the seven combinations and the other based on the median of the same seven models. The performance of the multi-model ensembles were evaluated on 12 meteorological stations, as well as for the entire domain of Ontario, using three temperature variables (average surface temperature, maximum surface temperature, and minimum surface temperature). Climate data for developing and validating the multi-model ensembles were collected from three major sources: the North American Coordinated Regional Downscaling Experiment, the Digital Archive of Canadian Climatological Data, and the Climactic Research Unit's TS v4.00 dataset. The results showed that the climate ensemble based on the mean generally outperformed the one based on the median, as well as each of the individual models. Future predictions under the Representative Concentration Pathway 4.5 (RCP4.5) scenario were generated using the multi-model ensemble based on the mean. This study provides credible and useful information for climate change mitigation and adaptation in Ontario.

Keywords Regional climate model · NA-CORDEX · Multi-model ensemble · Temperature variability · Ontario

Introduction

Climate models are complex mathematical representations of the climate system. They are based on established physical laws such as conservation of mass, energy and momentum (IPCC 2013). These models enable scientists to make future climate predictions, and aid in the analysis of climate impacts (Herrmann et al. 2016; Li et al. 2016; Wagner et al. 2017). Although climate models show significant errors in certain processes (such as cloud formation, leading to potential inaccuracies in timing and magnitude), they can provide plausible estimates for future variations in climate,

particularly for larger scales (Huo and Li 2013; Ragone et al. 2016; IPCC 2013).

To support climate change mitigation and adaptation, many climate modeling studies were conducted to analyze the climatic changes over Ontario, Canada. For instance, Wotton et al. (2003) used the Canadian Climate Centre coupled Global Climate Model (GCM) and the Hadley Centre's HadCM3 GCM to predict future changes in temperature and precipitation in Ontario. Wang et al. (2013) developed a statistical downscaling tool based on a stepwise cluster analysis method to obtain high-resolution climate projections for Ontario. Wang et al. (2015) developed a coupled dynamical-statistical approach to generate future climate projections for Ontario. However, many of the previous studies over the domain of Ontario were based mainly on a single climate model. To the authors' knowledge, there are no studies evaluating the performance of multi-model ensembles in Ontario.

An ensemble of multiple models is expected to outperform individual members of the ensemble (Barfus and Bernhofer 2014; Palmer et al. 2005). Multi-model ensembles are generally found to have a better performance than individual

Aly Al Samouly and Chanh Nien Luong are joint first authors.

✉ Zhong Li
zoeli@mcmaster.ca

¹ Department of Civil Engineering, McMaster University, Hamilton, ON L8S 4L7, Canada

² Department of Computing and Software Engineering, McMaster University, Hamilton, ON L8S 4L7, Canada

models. Ensembles are groupings of distinct models that undergo basic statistical calculations to produce a model that is more representative of the domain (Yan et al. 2016). In its simplest approach, multi-model ensembles are formed by merging a number of models with equal weights (Hagedorn et al. 2005; Jarsjo et al. 2017; Wallach et al. 2016). As suggested by Hagedorn et al. (2005), the simple multi-model ensemble has a higher likelihood for a better score than any single model. The most common climate models used are GCMs, to perform simulations across the globe (Zhai et al. 2018). Their largest drawback is their low resolution (Zhao et al. 2015). Regional Climate Models (RCMs) are tools used to downscale the GCMs to have a higher resolution in a relatively smaller area.

The objective of this study is to create, validate and evaluate the performance of a multi-model ensemble over the domain of Ontario, Canada. This entails the following: (1) evaluating the performance of seven GCM and RCM combinations for simulating the average and extreme temperatures across Ontario; (2) creating the multi-model mean and multi-model median ensembles and testing them over 12 stations and over the domain of Ontario; and (3) selecting an ensemble with the best performance and generating future climate predictions for Ontario using the selected multi-model ensemble. The results observed and gathered from this study are beneficial in the processes of decision-making and impact analyses for the future.

Methodology

In this study, two climate ensembles were developed based on four GCMs and four RCMs for Ontario, Canada. The individual GCM/RCM models as well as the developed

multi-model ensembles were evaluated by comparing model results with observational data of the present climate. The methodologies used for ensemble development and model validation are described as follows.

Study area and data collection

Ontario, the second largest province of Canada in area, covers approximately 1.076 million km² (Perera et al. 2000). It has a large number of water bodies, such as lakes and rivers, which significantly affect the climate in the region. The climate in Ontario can typically be considered as humid continental, except for parts of Northern Ontario under the influence of Hudson's Bay, which have a more maritime climate (Perera et al. 2000).

In this study, two sets of historical climate data were collected to validate the performance of the developed multi-model ensembles: observed data at 12 stations and Climatic Research Unit (CRU) gridded data. Twelve stations within Ontario, shown in Table 1 and Fig. 1, were chosen to validate the simulated data. The observed data were downloaded from the Digital Archive of Canadian Climatological Data provided by Environment and Climate Change Canada. Data with respect to three climate variables were downloaded: average surface temperature (*tas*), maximum surface temperature (*tasmax*) and minimum surface temperature (*tasmin*). The time period for each station varied between 1937 and 2015, depending on when observations began and stopped being recorded at the station. The second set of data based on observations was the CRU TS v4.00 dataset (Harris et al. 2014). The CRU TS v4.00 dataset was mainly developed by the UK's Natural Environment Research Council (NERC) and the US Department of Energy. It is maintained by the UK National Centre for Atmospheric Science (NCAS).

Table 1 Twelve selected stations

No.	Station name	Short name	Latitude	Longitude	Elevation	Observation data time period
1	Windsor Airport	Windsor	42.28°N	82.96°W	189.6 m	1940–2014
2	London International Airport	London	43.03°N	81.15°W	278.0 m	1940–2006
3	Toronto City Center Airport	Toronto City	43.63°N	79.40°W	173.4 m	1957–2006
4	Toronto Pearson International Airport	Toronto Pearson	43.68°N	79.63°W	76.8 m	1937–2015
5	Warton Airport	Warton	44.75°N	81.11°W	114.0 m	1947–2014
6	Ottawa International Airport	Ottawa	45.32°N	75.67°W	222.2 m	1938–2011
7	North Bay Airport	North Bay	46.36°N	79.42°W	370.3 m	1939–2014
8	Sault Ste Marie Airport	Sault Ste	46.48°N	84.51° W	192.0 m	1945–2012
9	Timmins Victor Power Airport	Timmins	48.57°N	81.38°W	383.4 m	1955–2011
10	Sioux Outlook Airport	Sioux	50.12°N	91.90°W	294.7 m	1938–2013
11	Moosonee	Moosonee	51.27°N	80.65°W	9.1 m	1932–2006
12	Big Trout Lake	Big Trout	53.83°N	89.87°W	224.1 m	1953–1992



Fig. 1 Location of the selected stations

Three variables were also downloaded for this dataset, *tmp*, *tmpmx* and *tmpmn*. These variables correspond to *tas*, *tasmax* and *tasmin*, respectively. These data are in a resolution of 0.5° and covered a time period from 1901 to 2015.

Multi-model ensemble

The simulated data were downloaded from the North American Coordinated Regional Downscaling Experiment (NA-CORDEX) archive, a branch of the International CORDEX Initiative (Giorgi et al. 2009; Lucas-Picher et al. 2013). The NA-CORDEX models provide large datasets of projections for historical and future predictions. These individual models can generate large uncertainties. This is evident in the comparison of the historical and observed data for each model. Data with respect to three climate variables and two scenarios were downloaded. The variables were *tas*, *tasmax* and *tasmin*. They were downloaded with respect to the historical (hist) and Representative Concentration Pathway 4.5 (RCP4.5) scenarios. Seven GCM and RCM combinations (shown in Table 2) were used in the analysis. The historical scenario covered a period between 1950 and 2005, while the RCP4.5 scenario covered a period between 2006 and 2099.

Multi-model ensembles are a logical and accepted technique in climate studies (Weigel et al. 2010). The approach taken in this study is to give every member of the ensemble equal weights. This method was used to develop two multi-model ensembles. The first ensemble developed was the multi-model mean ensemble. This ensemble was

Table 2 GCM and RCM Combinations

Scenario	GCM	RCM	Grid	Modeling institution	
hist	CanESM2	CanRCM4	0.44°	CCCma	
		CRCM5	0.44°	UQAM	
		RCA4	0.44°	SMHI	
	EC-EARTH	HIRHAM5	0.44°	DMI	
		RCA4	0.44°	SMHI	
		MPI-ESM-LR	CRCM5	0.44°	UQAM
		MPI-ESM-MR	CRCM5	0.44°	UQAM
RCP4.5	CanESM2	CanRCM4	0.44°	CCCma	
		CRCM5	0.44°	UQAM	
		RCA4	0.44°	SMHI	
	EC-EARTH	HIRHAM5	0.44°	DMI	
		RCA4	0.44°	SMHI	
		MPI-ESM-LR	CRCM5	0.44°	UQAM
		MPI-ESM-MR	CRCM5	0.44°	UQAM

created by taking the average temperature of each climate model at each corresponding time interval in the period between 1951 and 2005. The second model developed was the multi-model median ensemble. This was produced using a similar method to the multi-model mean ensemble. However, instead of taking the average of all models in a time period, the median was found. This was also an indicator of how much of an effect outliers in the simulations could have on the results. When comparing the simulated

data to the observed station data, the time period of the simulated data was matched with that of the observed data.

Evaluation of model performance

Numerous statistical methods were used to validate the CORDEX climate models and multi-model ensembles. Firstly, the coefficient of determination (R^2) was used as a measure of success of predicting the observed data from the simulated data. The coefficient of determination is characterized as proportion of variance ‘explained’ by the model (Nagelkerke 1991). R^2 ranges from 0 to 1. An R^2 value of 1 indicated that the simulated data perfectly fit the observation data. Secondly, the root mean square error (RMSE) was calculated. The RMSE has been used as a standard statistical metric to evaluate climate model performance (Chai and Draxler 2014). Thirdly, Taylor diagrams were constructed to provide a statistical summary of the correlation, RMSE and standard deviation of the comparison of each model, and of the multi-model ensemble, with the observed data. Fourthly, a monthly hypothesis test was carried out for each multi-model ensemble at each station, resulting in 144 hypothesis tests for each variable with the following null and alternate hypotheses (Katz 1992):

$$H_0 : \mu_1 - \mu_2 = 0, \quad (1)$$

$$H_a : \mu_1 - \mu_2 \neq 0, \quad (2)$$

where H_0 represents the null hypothesis and H_a represents the alternate hypothesis. μ_1 is the mean of the observed data and μ_2 is the mean of the simulated data. These sample means were calculated by finding the monthly average over the period of years at the station, then the hypothesis test was carried out. The confidence interval at a 90 and 95% confidence level was also found, which can be an alternative to formal hypothesis testing procedures (Katz 1992). The advantage of a confidence interval is that it gives information about the differences in the range of magnitudes, rather than answering ‘yes’ or ‘no’ to a hypothesis test. In the other words, the difference between observed and simulated data has a 90% or 95% chance of being contained in that range. Finally, monthly bias maps were plotted for each climate variable over the domain of Ontario to evaluate the performance of each model and the ensembles. Maps for the difference between the two temperature extremes, *tasmax* and *tasmin* were also plotted. After evaluating the performance of the two ensembles, two maps based on annual *tas* averages of the multi-model mean ensemble were plotted for the RCP4.5 scenario over two thirty year periods: 2040–2069 and 2070–2099. This provided a prediction for the trend of temperatures until 2099.

Results

Validation of model performance with observation data at selected stations

Twelve stations were selected within Ontario. The historical simulated *tas*, *tasmax* and *tasmin* were extracted and compared with the observed *tas*, *tasmax* and *tasmin* from each station. The data were then statistically evaluated to validate the models. Seven GCM and RCM combinations were analyzed at each station, and the multi-model mean and median ensembles were generated. The multi-model ensembles were graphed (shown in Fig. 2) and compared to the observed data using four different statistical techniques.

R^2 and RMSE values were calculated for each of the seven GCM and RCM combinations, as well as for the multi-model ensembles. The R^2 and RMSE values for the Toronto Pearson station are given in Table 3 as an example. The R^2 values compared each of the models to the observed data from the stations. The R^2 values for the two multi-model ensembles were highest in every station under every variable, indicating that they have the highest correlation to the observed data. Furthermore, the R^2 value from the multi-model mean ensemble was higher than that of the multi-model median ensemble. This demonstrates that the data used in creating the multi-model mean ensemble were not largely affected by outliers and have a stronger positive relationship with the observed data than any other model.

The RMSE results were similar to R^2 . The multi-model mean ensemble had the lowest RMSE values at all 12 stations for the *tas* variable. Wiarton was the only station where the RMSE of the multi-model mean ensemble was not lowest in *tasmax*. Four stations did not have the lowest RMSE in *tasmin*: these stations are Wiarton, Toronto City, Sault Ste Marie and Ottawa. In each instance where RMSE was not lowest in *tasmax* and *tasmin*, it was second lowest. This is mainly due to one model having a large RMSE (such as EC-EARTH - HIRHAM5 under *tasmax* in Wiarton), which distorts the multi-model mean ensemble RMSE. Similar to the R^2 results, the multi-model mean ensemble produced lower RMSE values than the multi-model median ensemble, indicating that the multi-model mean ensemble is a more accurate representation of the observed data. While the R^2 data show that there is a strong relationship between the multi-model mean ensemble and observed data, the RMSE shows that the multi-model mean ensemble is usually the best predictive model of the observation data.

Taylor diagrams (Taylor 2001) were constructed to help with the validation process. The monthly R^2 , RMSE and

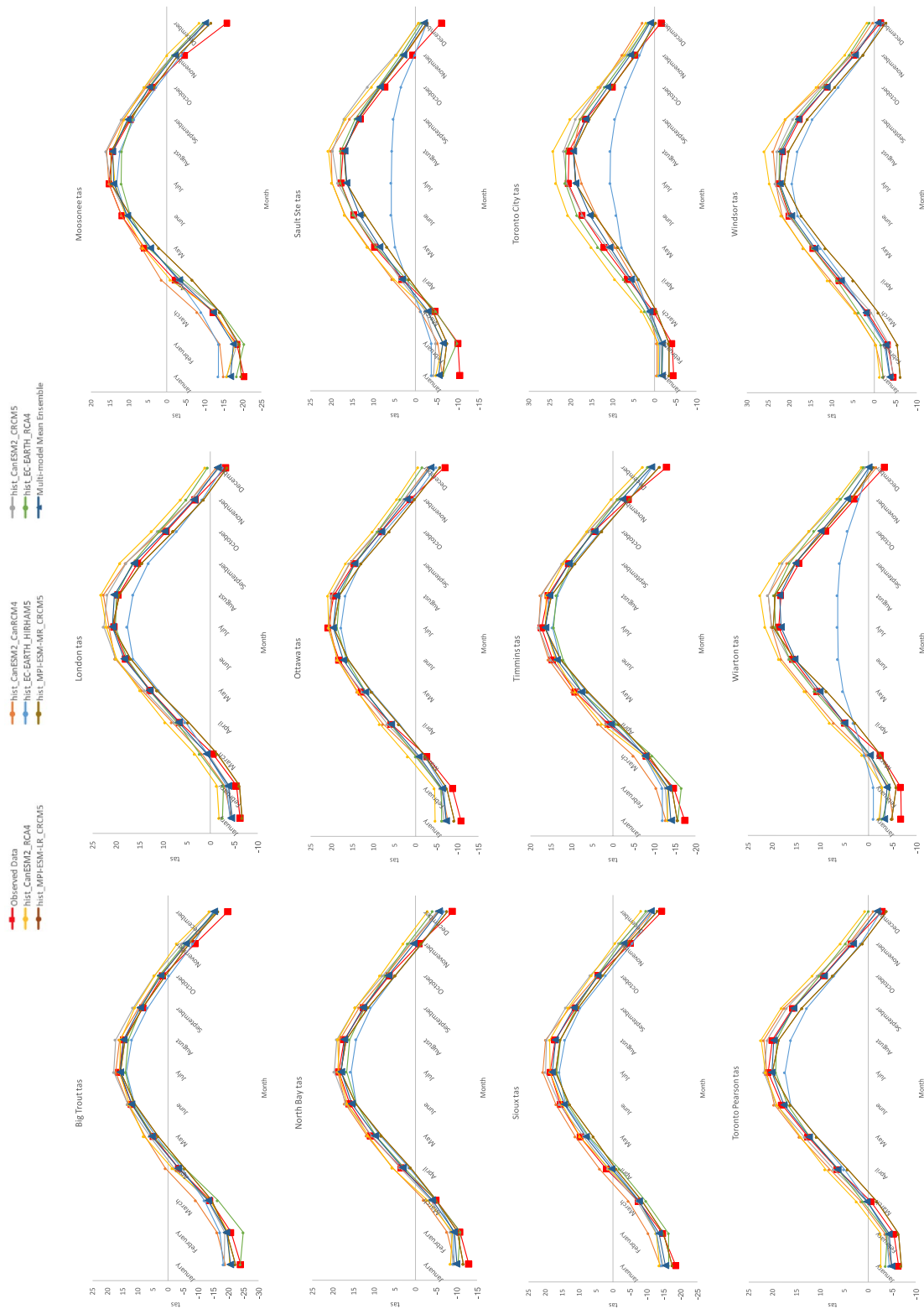


Fig. 2 Annual temperature variation

Table 3 R² and RMSE values for Toronto Pearson

Model		R ²			RMSE		
GCM	RCM	<i>tas</i>	<i>tasmax</i>	<i>tasmin</i>	<i>tas</i>	<i>tasmax</i>	<i>tasmin</i>
CanESM2	CanRCM4	0.92	0.92	0.89	9.95	12.77	13.20
CanESM2	CRCM5	0.93	0.93	0.91	7.84	8.58	10.13
CanESM2	RCA4	0.92	0.92	0.90	12.83	10.86	21.41
EC-EARTH	HIRHAM5	0.91	0.91	0.90	11.64	17.81	9.80
EC-EARTH	RCA4	0.91	0.92	0.88	9.52	10.21	13.61
MPI_ESM_LR	CRCM5	0.92	0.91	0.90	10.13	12.41	9.47
MPI_ESM_MR	CRCM5	0.92	0.91	0.90	10.13	12.41	9.47
Multi-Model Mean Ensemble		0.95	0.95	0.94	4.99	5.65	6.49

standard deviation were graphed (Fig. 3). By observing the Taylor diagram, we can see that the point corresponding to the multi-model mean ensemble (denoted as Average) is closest to the reference point in all the monthly Taylor diagrams. This indicates that the multi-model mean ensemble has the highest R² and lowest RMSE. The Taylor diagrams also show that the multi-model mean ensemble is more reliable and resembles the observed data much better than the multi-model median ensemble. This resulted in the rest of the analysis being done using only the multi-model mean ensemble.

Hypothesis tests were used to validate the data at 12 stations. The *p* values of the 144 hypothesis tests for *tas*, *tasmax* and *tasmin* were calculated. Since most of the 144 *p* values were below the significance level of 0.05 (90 for *tas*, 103 for *tasmax* and 122 for *tasmin*), the null hypothesis was rejected in most of the tests. For the tests where the null hypothesis was accepted (mainly in spring, summer, and fall months for *tas*; spring and summer months for *tasmin*; spring and fall months in *tasmax*), it is implied that the differences between observed and simulated data are not statistically significant. The results are consistent with the R² and RMSE results that have been discussed above. However, this cannot ensure that the data have been validated by this method, since the sample sizes are relatively small (about 40–50 depending on the period at different stations). It is possible that if the sample size grows larger, the null hypothesis may be rejected (Katz 1992). On the other hand, for the tests where the null hypothesis was rejected, which typically occurred in the winter periods or at stations located near large water bodies (for example, Sault Ste Marie or Toronto City), the differences between the observed and simulated values are statistically significant. To cope with the limitations of this validation approach and further validate the developed model, a confidence interval approach which is one alternative to formal hypothesis testing, was used.

The differences between observed and simulated data were also found to calculate the error ranges around these numbers. For example, at Big Trout, the average difference

in *tas* in January is -3.38 , while the confidence intervals for 90 and 95% confidence levels are $(-4.58, -2.19)$ and $(-4.38, -2.39)$, respectively. This is slightly different than the RMSE method, since the main purpose of the RMSE method is finding a positive number, which indicates the relationship between observed and simulated data, where the differences are squared. The maximum positive and maximum negative differences were also calculated. For *tas*, the maximum positive difference is 1.82 °C, and occurs in June at Toronto City, while the maximum negative difference is -5.59 °C, and occurs in December at Moosonee. In both *tasmax* and *tasmin*, the maximum positive difference occurs at Sault Ste, 4.23 °C in June for *tasmax*, and 0.45 °C in May for *tasmin*, while the maximum negative difference occurs in the winter period at Moosonee, -5.59 °C in December for *tasmax*, and -7.31 °C in January for *tasmin*. Since the models could not be validated through comparison of the means, confidence intervals were used for further model validation.

The confidence interval approach allows the approximation of ranges where the differences between observed and simulated values fall into a certain confidence level (typically 90% or 95%). The confidence intervals could provide the differences between observed and simulated values at different stations in different months. The average, maximum positive and maximum negative differences between observed and simulated means were calculated. In *tas*, the average difference is -0.646 °C, while maximum positive and maximum negative differences are 1.82 and -5.59 °C, respectively. In *tasmax* and *tasmin*, the values are 0.21 , 4.23 , -3.87 °C and -2.06 , 0.45 , -7.31 °C, respectively. The confidence intervals suggest that the multi-model ensemble performs well in all months except for the winter season and is within an acceptable error range (Suklitsch et al. 2011). The calculated average confidence intervals, which were obtained by taking the average of absolute differences of all stations from January to December, were ± 1.27 , ± 1.41 , ± 2.16 °C for *tas*, *tasmax* and *tasmin*, respectively. When the winter months were omitted from the calculation, the average confidence intervals obtained by taking the

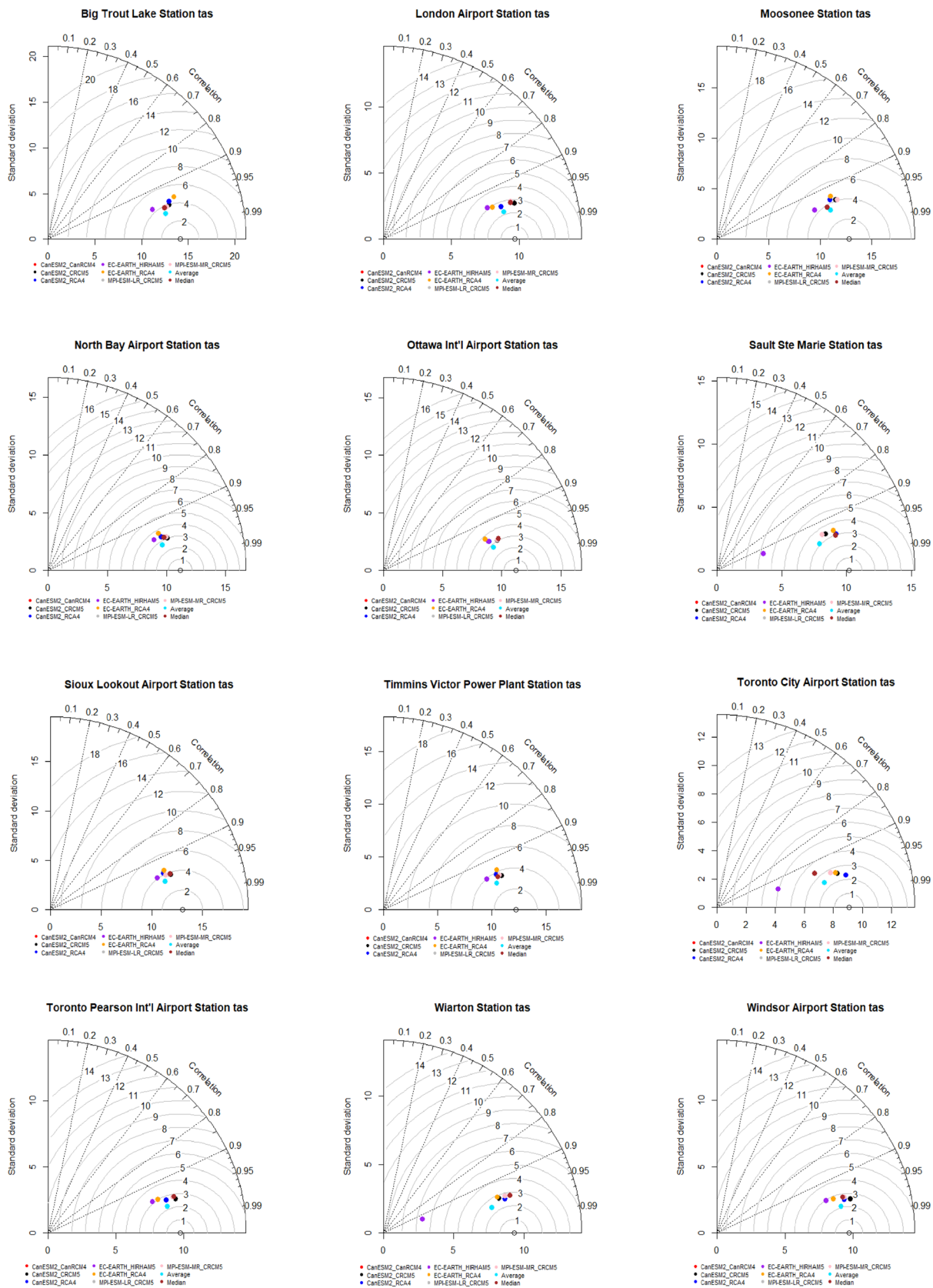


Fig. 3 Taylor diagrams for 12 stations—*tas*

average of absolute differences of all stations from January to December, were ± 0.77 , ± 1.27 , ± 1.46 °C for *tas*, *tasmax* and *tasmin*, respectively. Moreover, the largest confidence intervals were found in *tas*, *tasmax* and *tasmin* at 90 and 95% confidence levels. In *tasmax*, the largest interval occurred in December at Big Trout, which was (-4.55 , -2.46 °C) and (-4.76 , -2.25 °C) at the 90% and 95% confidence levels respectively. Similarly, in *tas* and *tasmin*, the largest intervals occurred in December at Moosonee, which were (-6.66 , -4.53 °C) and (-6.87 , -4.31 °C) for *tas*, (-8.39 , -6.01 °C) and (-8.62 , -5.77 °C) for *tasmin*.

Validation of model performance with observation-based gridded datasets

The grids from the gridded CRU observation data (Harris et al. 2014) and NA-CORDEX simulated data were each placed on a grid of the map of Ontario, and the closest point to each coordinate was found. This enables us to compare the temperatures at the specific points while minimizing errors. Results from the 12 stations strongly suggest that the multi-model mean ensemble resembles the observed data more accurately than the multi-model median ensemble. Analyses of the gridded data were only carried out using the seven models and the multi-model mean ensemble. Various techniques were used to validate the gridded datasets.

The R^2 values were generated by comparing the observed and simulated data across every point on the grid. The results are shown in Table 4. The R^2 values are highest for the multi-model mean ensemble under all three variables. These results show that the multi-model mean ensemble and observed data have a stronger positive relationship than any other model over the entire domain of Ontario.

The RMSE values were also generated to compare the observed and simulated data. The RMSE for the multi-model mean ensemble is the lowest under *tasmax*; however, it is the third lowest under *tas* and *tasmin*. This is due to some individual models (such as EC-EARTH - HIRHAM5 under *tasmax*) with very large RMSE values, which distort the mean.

The bias at each point on the grids was calculated by finding the difference between the observed and simulated data (Fig. 4). This showed locations and periods during the year at which the biases are large or small. The maps indicate that the simulated models show a larger bias in the winter months, generating values that are lower than observed. The models also consistently generate large biases near the Great Lakes and Hudson's Bay. The largest absolute bias was 11.20 °C at (-86.32° , 55.96°) which falls right off the shore of Northern Ontario, in Hudson's Bay.

Analysis of temperature variability over Ontario, Canada

The developed multi-model mean ensemble was used to predict future temperature variability over Ontario. The scenario analyzed was the Representative Concentration Pathway (RCP) 4.5. This scenario stabilizes a radiative forcing of 4.5 W/m² in 2100, allowing climate models to explore the climate system's response to moderating the anthropogenic components in radiative forcing (Thomson et al. 2011). Many studies choose to focus on the RCP4.5 scenario (Laprise et al. 2013; Lee and Wang 2014; Rotstayn et al. 2012). The multi-model mean ensemble predicts an increase in temperature of 2.89 °C between the fifty-year historical period of 1951–2005 and the thirty-year future prediction period of 2040–2069 (Fig. 5).

The future temperature projections developed in this study have been made available to the public and can be downloaded through the Mendeley data portal (<https://doi.org/10.17632/6dtpjschn7.3>). These datasets are the multi-model mean ensemble specific to *tas*, *tasmax*, and *tasmin* over the two periods of 2040–2069 and 2070–2099, covering the Province of Ontario using a grid of approximately 0.44°. This allows open access to high-resolution temperature projections for assessing climate change impacts on local communities and formulating effective mitigation and adaptation strategies for Ontario.

Table 4 R^2 values across the whole domain

Model		R^2			RMSE		
GCM	RCM	<i>tas</i>	<i>tasmax</i>	<i>tasmin</i>	<i>tas</i>	<i>tasmax</i>	<i>tasmin</i>
CanESM2	CanRCM4	0.945	0.946	0.928	16.339	13.321	28.020
CanESM2	CRCM5	0.977	0.969	0.963	6.735	5.679	15.253
CanESM2	RCA4	0.967	0.963	0.946	9.997	8.374	21.100
EC-EARTH	HIRHAM5	0.958	0.940	0.963	12.190	19.963	15.117
EC-EARTH	RCA4	0.953	0.950	0.933	7.678	8.718	12.217
MPI-ESM-LR	CRCM5	0.982	0.974	0.970	3.772	8.341	5.477
MPI-ESM-MR	CRCM5	0.982	0.974	0.970	3.772	8.341	5.477
Multi-Model Mean Ensemble		0.983	0.977	0.975	4.108	4.511	9.753

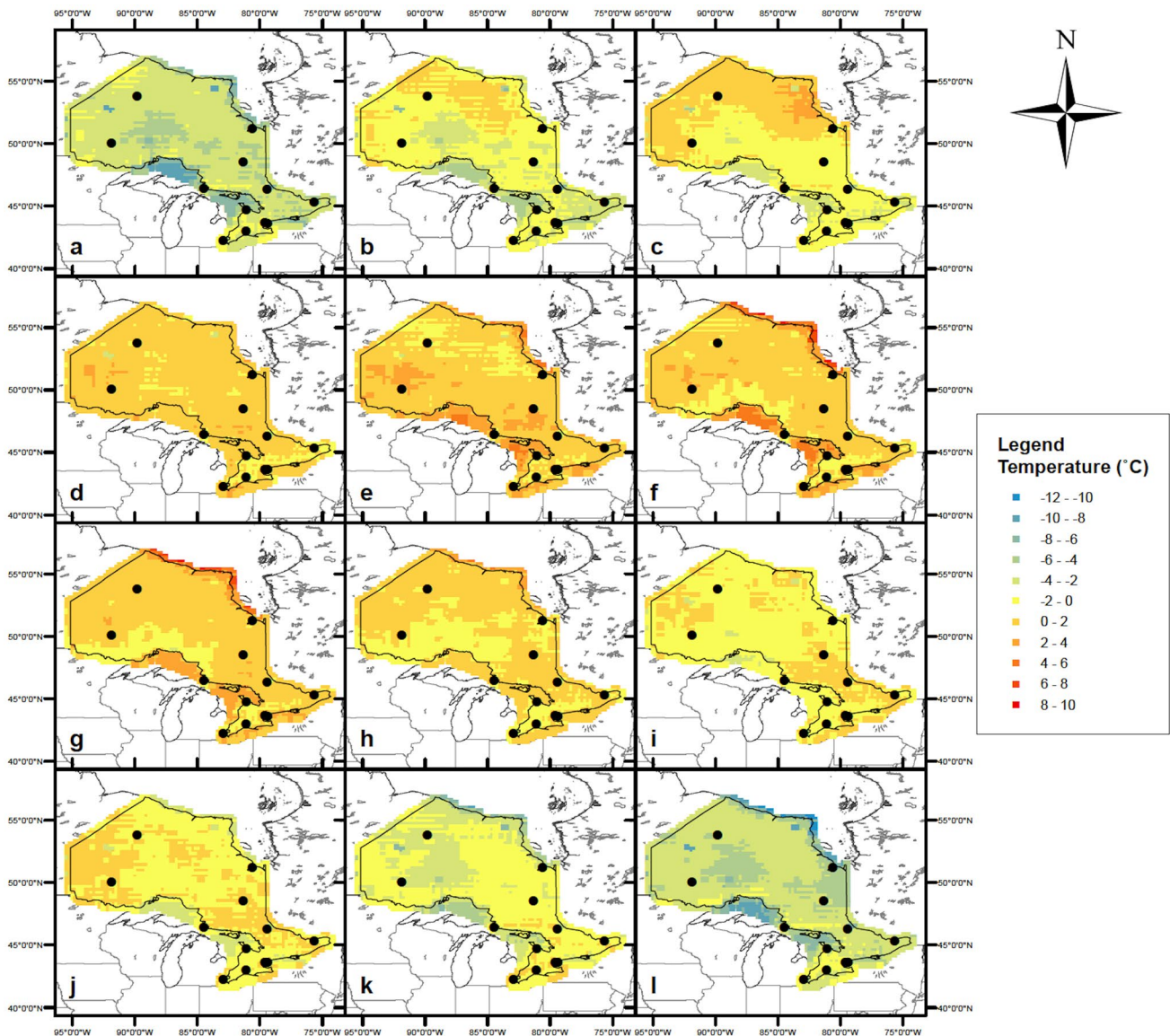


Fig. 4 Temperature bias: **a** January, **b** February, **c** March, **d** April, **e** May, **f** June, **g** July, **h** August, **i** September, **j** October, **k** November, **l** December

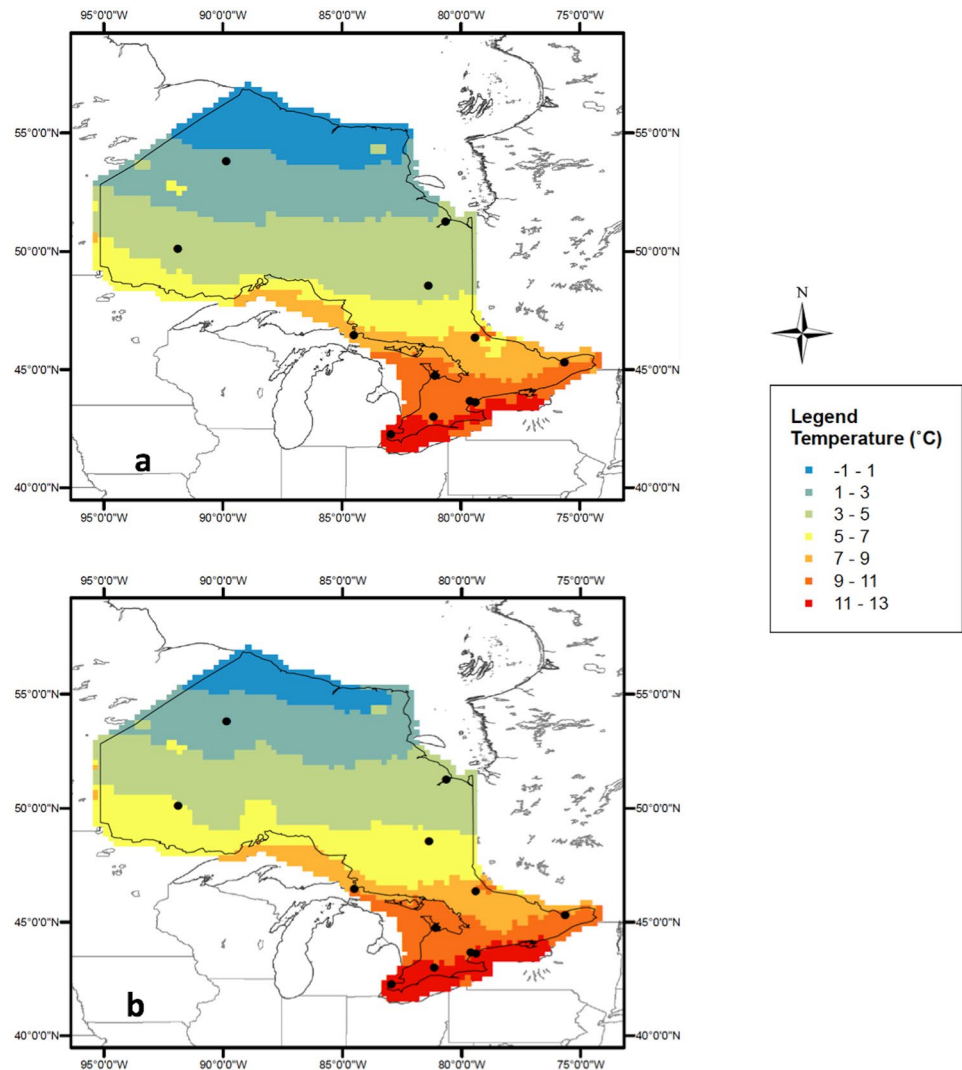
Discussion

The validation with observation data at the 12 selected stations shows that the data from the multi-model mean ensemble has a strong positive relationship with observation data and were insignificantly affected by outliers. These results are in agreement with results from Kirtman and Min (2009), where the multi-model ensemble has the highest correlation, indicating that the multi-model ensemble performs better than the individual models. The results are similar to those from Lambert and Boer (2001), who also state that the multi-model mean tends to have a lower RMSE than most individual models. Krishnamurti et al. (2000) also discovered an improvement in RMSE

for a when a multi-model ensemble was used over a single model.

Results from the hypothesis tests were also used to validate the model performance. Since most of the 144 *p* values were below the significance level of 0.05. According to Katz (1992), such statistically significant discrepancies do not necessarily suggest poor performance of the model. Devineni et al. (2008) used a similar approach where the null hypothesis was also rejected. To address the limitations of this approach, a confidence interval approach which is one alternative to formal hypothesis testing, was used. Results from the confidence interval indicated a larger variation in temperature in the winter months, similar to Dasari et al. (2014). Winter months have yielded larger biases and RMSE

Fig. 5 30-year annual mean temperature over Ontario under the RCP4.5 scenario: **a** 2040–2069 and **b** 2070–2099



values than winter months in previous studies (Rozante et al. 2014). Additionally, stations in locations near large water bodies such as Big Trout and Moosonee yielded large biases. These observations are in agreement with Xue et al. (2017), which states that lake surface temperature and ice coverage have a large impact on regional climate.

The validation with observation-based gridded datasets shows that the multi-model mean ensemble generated R^2 values larger than any individual model in all three climate variables. This is not uncommon as seen by studies done by Zhang et al. (2011) and Kirtman and Min (2009). Similarly, RMSE for the multi-model mean ensemble is the lowest for all climate variables except *tasmax*. The EC-EARTH - HIRHAM5 has a large RMSE value that distorts the mean. This model has shown large RMSE under maximum temperatures in the past (Mezghani et al. 2017). To find the locations and times where large deviations from observed data occur, the bias at every coordinate was calculated throughout the year. The results show large biases in the winter months.

Climate models are notorious for having a relatively poor performance during winter months due to deficiencies in model surface physics (Dasari et al. 2014). Furthermore, large biases were observed near large water bodies. One of the main unresolved issues in climate modelling is the erroneous representation of the lake-ice-atmosphere interaction in RCMs (Xue et al. 2017), which is a reason for this large uncertainty.

According to the ensemble outputs, the temperature would increase by 2.89 °C on average from the period of 1951–2015 towards 2040–2069. This falls within the range of the assumptions of the Government of Ontario, who stated an increase in temperatures in the range of 2.50–3.70 °C by 2050 (MOECC 2011). This rise in temperatures would have adverse effects on many natural resources in Ontario (MOECC 2011). For example, water levels are expected to decrease due to a reduction in precipitation and increased demand. An increase in extreme weather events is expected to have direct effects on the energy infrastructure (MOECC

2011). Increasing temperatures are also predicted to increase costs by \$5 billion per year in 2020 to between \$21 billion and \$43 billion per year by the 2050s (Demerse 2016).

Conclusions

Climate models that were accessed through the NA-CORDEX were used in creating two multi-model ensembles, a multi-model mean ensemble and a multi-model median ensemble. Two observation-based datasets were used to validate the performance of the developed ensembles. The models were tested over 12 stations in Ontario using data obtained from the Government of Canada, and the ensembles were also compared over the whole domain of Ontario using the CRU TS v4.00 observationally based gridded data.

The multi-model mean ensemble was found to be the superior model at all 12 stations. This was evident when the evaluation for all seven models as well as the two multi-model ensembles was carried out. This was an indicator that the multi-model mean ensemble was more accurate across the domain of Ontario. The rest of the study was completed using the multi-model mean ensemble and the results proved that it is preferable to the multi-model median ensemble as well as every individual model. The future predictions using the multi-model ensemble showed a heating trend in all three variables analyzed, *tas*, *tasmax* and *tasmin*. This is in agreement with the trends from the International Panel of Climate Change's Fifth Assessment Report.

By far, the largest limitation in this study was the significant difference in the nature of the models used in the ensemble. These models are likely to have a different number of variables which is the reason that some are more representative of the observed data. This would create complications in the statistical methods used to create the ensembles. Knowing how many variables are representative of each model would help in adding different weights to specific models when creating the multi-model ensembles. A similar method was used by (Tebaldi and Knutti 2007); however, their model weights were based on bias and convergence criteria.

Acknowledgements This research was supported by the Natural Science and Engineering Research Council of Canada. We acknowledge the World Climate Research Programme's Working Group on Regional Climate, and the Working Group on Coupled Modelling, former coordinating body of CORDEX and responsible panel for CMIP5. We also thank the climate modelling groups (listed in Table 2 of this paper) for producing and making available their model output. We also acknowledge the Earth System Grid Federation infrastructure an international effort led by the U.S. Department of Energy's Program for Climate Model Diagnosis and Intercomparison, the European Network for Earth System Modelling and other partners in the Global Organisation for Earth System Science Portals (GO-ESSP). We would like to express our very great appreciation to Dr. Alessandro Selvitella for his valuable

advice and guidance for the statistical techniques used in this research paper.

References

- Barfus K, Bernhofer C (2014) Assessment of GCM performances for the Arabian Peninsula, Brazil, and Ukraine and indications of regional climate change. *Environ Earth Sci* 72:4689–4703. <https://doi.org/10.1007/s12665-014-3147-3>
- Chai T, Draxler RR (2014) Root mean square error (RMSE) or mean absolute error (MAE)?—arguments against avoiding RMSE in the literature. *Geosci Model Dev* 7:1247–1250. <https://doi.org/10.5194/gmd-7-1247-2014>
- Dasari HP, Salgado R, Perdigao J, Challa VS (2014) A regional climate simulation study using WRF-ARW model over Europe and evaluation for extreme temperature weather events International. *J Atmos Sci* 704079:1–22
- Demerse C (2016) Ignoring climate change will cost us too—big time. Clean Energy Canada. <http://cleanenergycanada.org/ignoring-climate-change-will-cost-us-too-big-time/>. Accessed 22 Sep 2017
- Devineni N, Sankarasubramanian A, Ghosh S (2008) Multimodel ensembles of streamflow forecasts: role of predictor state in developing optimal combinations. *Water Resour Res* 44:W09404. <https://doi.org/10.1029/2006WR005855>
- Giorgi F, Jones C, Asrar GR (2009) Addressing climate information needs at the regional level: the CORDEX framework. *World Meteorol Organ (WMO) Bull* 58:175
- Hagedorn R, Doblas-Reyes FJ, Palmer TN (2005) The rationale behind the success of multi-model ensembles in seasonal forecasting—I. Basic concept. *Tellus A* 57:219–233. <https://doi.org/10.1111/j.1600-0870.2005.00103.x>
- Harris I, Jones PD, Osborn TJ, Lister DH (2014) Updated high-resolution grids of monthly climatic observations—the CRU TS3.10 dataset. *Int J Climatol* 34:623–642. <https://doi.org/10.1002/joc.3711>
- Herrmann F, Kunkel R, Ostermann U, Vereecken H, Wendland F (2016) Projected impact of climate change on irrigation needs and groundwater resources in the metropolitan area of Hamburg (Germany) *Environ Earth Sci* 75 <https://doi.org/10.1007/s12665-016-5904-y>
- Huo AD, Li H (2013) Assessment of climate change impact on the stream-flow in a typical debris flow watershed of Jianzhuangcuan catchment in Shaanxi Province. *China Environ Earth Sci* 69:1931–1938. <https://doi.org/10.1007/s12665-012-2025-0>
- IPCC (2013) Climate change 2013: The physical science basis. In: Stocker TF, Qin D, Plattner G-K, Tignor M, Allen SK, Boschung J, Nauels A, Xia Y, Bex V, Midgley PM (eds) Contribution of working group I to the fifth assessment report of the intergovernmental panel on climate change. Cambridge University Press, Cambridge, pp 1535
- Jarsjo J, Tornqvist R, Su Y (2017) Climate-driven change of nitrogen retention-attenuation near irrigated fields: multi-model projections for Central Asia. *Environ Earth Sci* 76 <https://doi.org/10.1007/s12665-017-6418-y>
- Katz RW (1992) Role of statistics in the validation of general circulation models. *Clim Res* 2:35–45
- Kirtman BP, Min D (2009) Multimodel ensemble ENSO prediction with CCSM and CFS. *Mon Weather Rev* 137:2908–2930
- Krishnamurti TN et al (2000) Multimodel ensemble forecasts for weather and seasonal climate. *J Clim* 13:4196–4216
- Lambert SJ, Boer GJ (2001) CMIP1 evaluation and intercomparison of coupled climate models. *Clim Dyn* 17:83–106
- Laprise R et al (2013) Climate projections over CORDEX Africa domain using the fifth-generation Canadian Regional

- Climate Model (CRCM5). *Clim Dyn* 41:3219–3246. <https://doi.org/10.1007/s00382-012-1651-2>
- Lee JY, Wang B (2014) Future change of global monsoon in the CMIP5. *Climate Dynamics* 42:101–119. <https://doi.org/10.1007/s00382-012-1564-0>
- Li Z, Huang G, Wang X, Han J, Fan Y (2016) Impacts of future climate change on river discharge based on hydrological inference: a case study of the Grand River Watershed in Ontario. *CanSci Tot Environ* 548:198–210. <https://doi.org/10.1016/j.scitotenv.2016.01.002>
- Lucas-Picher P, Somot S, Deque M, Decharme B, Alias A (2013) Evaluation of the regional climate model ALADIN to simulate the climate over North America in the CORDEX framework. *Clim Dyn* 41:1117–1137. <https://doi.org/10.1007/s00382-012-1613-8>
- Mezghani A et al (2017) CHASE-PL Climate Projection dataset over Poland—Bias adjustment of EURO-CORDEX simulations. *Earth Syst Sci Data Discuss* 2017:1–29. <https://doi.org/10.5194/essd-2017-51>
- MOECC (2011) Climate Ready: Ontario's Adaptation Strategy and Action Plan 2011–2014. Ontario Ministry of the Environment and Climate Change, Canada
- Nagelkerke NJD (1991) A note on a general definition of the coefficient of determination. *Biometrika* 78:691–692. <https://doi.org/10.1093/biomet/78.3.691>
- Palmer TN, Doblus-Reyes FJ, Hagedorn R, Weisheimer A (2005) Probabilistic prediction of climate using multi-model ensembles: from basics to applications. *Philos Trans R Soc B* 360:1991–1998
- Perera AH, Euler D, Thompson ID (2000) Ecology of a managed terrestrial landscape: patterns and processes of forest landscapes in Ontario. UBC Press in cooperation with the Ontario Ministry of Natural Resources, Vancouver
- Ragone F, Lucarini V, Lunkeit F (2016) A new framework for climate sensitivity and prediction: a modelling perspective. *Clim Dyn* 46:1459–1471. <https://doi.org/10.1007/s00382-015-2657-3>
- Rotstayn LD, Jeffrey SJ, Collier MA, Dravitzki SM, Hirst AC, Syktus JJ, Wong KK (2012) Aerosol- and greenhouse gas-induced changes in summer rainfall and circulation in the Australasian region: a study using single-forcing climate simulations. *Atmos Chem Phys* 12:6377–6404. <https://doi.org/10.5194/acp-12-6377-2012>
- Rozante J, Moreira D, Godoy R, Fernandes A (2014) Multi-model ensemble: technique and validation. *Geosci Model Dev Discuss* 7:2933–2959
- Suklitsch M, Gobiet A, Truhetz H, Awan NK, Göttel H, Jacob D (2011) Error characteristics of high resolution regional climate models over the Alpine area. *Clim Dyn* 37:377–390
- Taylor KE (2001) Summarizing multiple aspects of model performance in a single diagram. *J Geophys Res Atmos* 106:7183–7192. <https://doi.org/10.1029/2000jd900719>
- Tebaldi C, Knutti R (2007) The use of the multi-model ensemble in probabilistic climate projections. *Philos Trans R Soc A* 365:2053–2075. <https://doi.org/10.1098/rsta.2007.2076>
- Thomson AM et al (2011) RCP4. 5: a pathway for stabilization of radiative forcing by 2100. *Clim Change* 109:77
- Wagner T, Themessl M, Schuppel A, Gobiet A, Stigler H, Birk S (2017) Impacts of climate change on stream flow and hydro power generation in the Alpine region *Environ Earth Sci*. <https://doi.org/10.1007/s12665-016-6318-6>
- Wallach D, Mearns L, Ruane A, Rotter R, Asseng S (2016) Lessons from climate modeling on the design and use of ensembles for crop modeling. *Clim Change* 139:551–564. <https://doi.org/10.1007/s10584-016-1803-1>
- Wang X et al (2013) A stepwise cluster analysis approach for down-scaled climate projection—a Canadian case study. *Environ Model Softw* 49:141–151
- Wang XQ, Huang GH, Lin QG, Nie XH, Liu JL (2015) High-resolution temperature and precipitation projections over Ontario, Canada: a coupled dynamical-statistical approach. *Quart J R Meteorol Soc* 141:1137–1146
- Weigel AP, Knutti R, Liniger MA, Appenzeller C (2010) Risks of model weighting in multimodel climate projections. *J Clim* 23:4175–4191. <https://doi.org/10.1175/2010jcli3594.1>
- Wotton B, Martell D, Logan K (2003) Climate change and people-caused forest fire occurrence in Ontario. *Clim Change* 60:275–295
- Xue PF, Pal JS, Ye XY, Lenters JD, Huang CF, Chu PY (2017) Improving the simulation of large lakes in regional climate modeling: two-way lake–atmosphere coupling with a 3D hydrodynamic model of the great lakes. *J Clim* 30:1605–1627. <https://doi.org/10.1175/Jcli-D-16-0225.1>
- Yan RH, Gao JF, Li LL (2016) Streamflow response to future climate and land use changes in Xinjiang basin, China. *Environ Earth Sci* 75 <https://doi.org/10.1007/s12665-016-5805-0>
- Zhai Y, Huang G, Wang X, Zhou X, Lu C, Li Z (2018) Future projections of temperature changes in Ottawa, Canada through stepwise clustered downscaling of multiple GCMs under RCPs. *Clim Dyn*. <https://doi.org/10.1007/s00382-018-4340-y>
- Zhang Q, Dool H, Saha S, Mendez M, Becker E, Peng P, Huang J (2011) Preliminary evaluation of multi-model ensemble system for monthly and seasonal prediction. In: 36th NOAA annual climate diagnostics and prediction workshop, Fort Worth, USA, 3–6 October 2011. Science and Technology Infusion Climate Bulletin, pp 124–131
- Zhao N, Chen CF, Zhou X, Yue TX (2015) A comparison of two downscaling methods for precipitation in China. *Environ Earth Sci* 74:6563–6569. <https://doi.org/10.1007/s12665-015-4750-7>