Chances of Short-Term Cooling Estimated from a Selection of CMIP5-Based Climate Scenarios during 2006–35 over Canada

PATRICK GRENIER AND RAMÓN DE ELÍA
Ouranos, and Centre pour l’Étude et la Simulation du Climat à l’Échelle Régionale, Université du Québec à Montréal, Montréal, Québec, Canada

DIANE CHAUMONT
Ouranos, Montréal, Québec, Canada

(Manuscript received 24 March 2014, in final form 8 January 2015)

ABSTRACT

The path toward a warmer global climate is not smooth, but, rather, is made up of a succession of positive and negative temperature trends, with cooling having more chance to occur the shorter the time scale considered. In this paper, estimates of the probabilities of short-term cooling ($P_{\text{cool}}$) during the period 2006–35 are performed for 5146 locations across Canada. Probabilities of cooling over durations from 5 to 25 yr come from an ensemble of 60 climate scenarios, based on three different methods using a gridded observational product and CMIP5 climate simulations. These methods treat interannual variability differently, and an analysis in hindcast mode suggests they are relatively reliable. Unsurprisingly, longer durations imply smaller $P_{\text{cool}}$ values; in the case of annual temperatures, the interdecile range of $P_{\text{cool}}$ values across Canada is, for example, $-2\%$–$18\%$ for 25 yr and $-40\%$–$46\%$ for 5 yr. Results vary slightly with the scenario design method, with similar geographical patterns emerging. With regards to seasonal influence, spring and winter are generally associated with higher $P_{\text{cool}}$ values. Geographical $P_{\text{cool}}$ patterns and their seasonality are explained in terms of the interannual variability over background trend ratio. This study emphasizes the importance of natural variability superimposed on anthropogenically forced long-term trends and the fact that regional and local short-term cooling trends are to be expected with nonnegligible probabilities.

1. Introduction

Climate scientists recognize that, although Earth’s surface temperatures have entered a period of general long-term increase, temporary stalling or cooling periods may occur at any spatial scale (Easterling and Wehner 2009; Knight et al. 2009; Meehl et al. 2011; de Elía et al. 2013; Deser et al. 2014). Most long-term trends for the twenty-first century are projected as positive with a high degree of confidence, because anthropogenic greenhouse gases are the dominant factor for that period and their radiative forcing is expected to increase (van Vuuren et al. 2011; Stocker et al. 2013). Short-term trend distributions are tied to this global forcing as well but are also expected to reflect the fast-changing influence of secondary factors like natural forcings (e.g., volcanic aerosols and solar cycles) and unforced natural variability [e.g., the North Atlantic Oscillation (NAO) and the El Niño–Southern Oscillation (ENSO)].

The recent period (~1998–2012) offers an example of temporary warming slowdown on the global scale (Morice et al. 2012). This could have resulted from naturally prevailing La Niña–like conditions in the tropical Pacific (Kosaka and Xie 2013) accompanied by subsurface ocean heat uptake higher than in decades that experienced more warming (Meehl et al. 2011; England et al. 2014; Risbey et al. 2014). The NAO could be implicated as well, at least for the Northern Hemisphere (Li et al. 2013). Other possible causes, involving a decrease in top-of-atmosphere radiative forcing, have been proposed: decreased ozone-depleting substances and methane emissions (Estrada et al. 2013), decreased solar irradiance during cycle 23’s declining phase (Lean 2010; Kopp and Lean 2011), increased sulfur emissions in Asia (Kaufmann et al. 2011), and declining stratospheric water vapor concentrations after 2000 (Solomon...
et al. 2010). The recent slowdown may be in part a measurement artifact due to incomplete global coverage (Cowtan and Way 2013), but in this case it may be only a matter of time before a 10- or 15-yr global cooling trend actually occurs.

At the regional and local scales, short-term cooling trends are expected to occur with higher probabilities, because interannual and interdecadal variability is generally higher at smaller spatial scales (Deser et al. 2012; Maraun 2013). For North America, Deser et al. (2014) have shown that internal variability can result in as much as ~40% of simulations generated by state-of-the-art global coupled models showing cooling trends over some regions during the period 2010–35.

Although the possibility of future cooling trends in a long-term warming path is undisputed, most studies on the subject are based on direct outputs from climate models, which present differences in many statistical properties when compared with observational products. For example, van Oldenborgh et al. (2013) have examined the Coupled Model Intercomparison Project phase 5 (CMIP5) multimodel ensemble and concluded that the distribution of surface temperature trends (per degree of global warming) for 1950–2011 and north of 45°S is underdispersive; over Canada, their analysis suggests an underforecasting bias, with observed trends almost all within the 20th–99th-percentile bracket of CMIP5 trends. In another study, Bhend and Whetton (2013) also concluded that there were inconsistencies between observed and CMIP3–CMIP5 trends, including for regions in Canada. Other relevant studies of temperature trends at different time and space scales in multimodel ensembles include those of Santer et al. (2011), Sakaguchi et al. (2012), Stott et al. (2013), and Knutson et al. (2013).

Reported inconsistencies between simulated and observed trends are not the only reason why computing future probability of cooling trend (Pcool) values based on climate model simulations is controversial (Curry 2011; Stephenson et al. 2012). Another important reason is that a “true” value of Pcool would have to be based on a perfectly representative distribution of trends, which multimodel ensembles cannot guarantee (von Storch and Zwiers 2013). Other issues related to probabilistic projections of climate include the difficulty of assessing reliability given a single real climate trajectory, inferring models’ future skill from past skill, subjectivity in model weighting, model tuning, and anthropogenic forcing uncertainty (Räisänen and Palmer 2001; Tebaldi and Knutti 2007; Yokohata et al. 2013). However, it may be argued that estimations of Pcool should be provided by climate experts when needed, with limitations clearly stated. The rationale is that otherwise, decisions are made “with implied assessments of relative likelihoods that depart, perhaps significantly, from experts’ best estimates” (Hall et al. 2005, p. 346). Notably, Räisänen and Palmer (2001), Schneider (2002), and Collins (2007) encouraged probabilistic climate forecasting.

This study’s main objective is to estimate Pcool for short-term durations (from 5 to 25 yr) over Canada during the current period (2006–35) in ways that reflect how numerical climate simulations are actually used by climate service institutions to provide decision makers with local climate scenarios (Huard et al. 2014). A local univariate climate scenario is defined here as a time series that describes a path one climate variable may plausibly follow over a very small area and during a given period of time. Plausibility requires statistical properties consistent with observations for the past segment, continuity at the past–future junction, and physical credibility for the future segment. Climate scenarios are designed by merging statistical properties from numerical model simulations and observational products, with the direct usage of simulations seen as one of the options. No single scenario is a prediction; however, as an ensemble, scenarios aim to cover the range that includes the real future climate trajectory.

A secondary objective is to evaluate whether climate scenario design methods other than the direct use of simulations reduce offsets between observational and simulated Pcool distributions during the calibration period (1962–2010). Two such alternative methods are used. One method performs statistical adjustment of the simulated interannual variability, while the other involves using a simple autoregressive model to generate variability.

It is important to emphasize local short-term cooling trends in the context of global warming for at least two reasons. First, some decisions do require information related to the climate’s short-term evolution or are based on changes from climate normal, and it may be important to know the range of temperature trends. Second, global warming and consequential environmental changes represent a scientific challenge but also a political issue, implying that it might be tempting for some interest groups to create and spread disinformation. For example, the recent warming slowdown is known to have fueled contrarian views, as discussed by Kaufmann et al. (2011) and Cowtan and Way (2013). Thus it must be clearly demonstrated that short-term cooling trends and long-term global warming represent two intertwined phenomena and that the occurrence of the former is in no way a refutation of the latter (Santer et al. 2011).

The paper is divided as follows. In section 2, datasets and methodological choices are described. Hindcast skill results for 1962–2010 and probabilistic forecasts for 2006–35 are presented in sections 3 and 4, respectively. Section 5 consists of a summary with discussion.
2. Data and methods

Numerical model simulations are often used directly as scenarios for making inferences about aspects of Earth’s real future climate (e.g., Deser et al. 2014). This is justified when statistical properties of interest are judged to be realistic. But a more general approach to climate scenario design would be to merge information from simulations and observations, with all the weight on simulations being just one option. The rationale for generalizing the approach is that there exist statistical offsets between simulations and local realizations of nature, and part of this is due to scale mismatch and the inherent imperfections of models. This causes discontinuities at the junction of the past-observed and future-simulated segments, among other problems. However, part of these offsets may also result from limitations in observational products (particularly in northern Canada, where the network of stations is sparse) and from the comparison period being too short for both observations and simulations to represent the full phase space of the local climate. For this and other reasons, there is no scientific consensus on the legitimacy of methods akin to bias correction (Thomeßl et al. 2012; Haerter et al. 2011; Ehret et al. 2012).

In this study, $P_{\text{cool}}$ uncertainty related to the choice of scenario design is addressed by considering three different methods. These are based on the same observational product and simulation ensemble but merge the statistical properties differently. The property of interest here is the short-term trend distribution. Because it is difficult to design climate scenarios in a way that directly ensures their short-term trends will be consistent with observations, methods (other than the direct use of simulations) address the problem indirectly by seeking agreement in interannual variability (standard deviation of residuals around a background state or long-term time-varying trend). Offsets in climate simulation variability have the potential to affect trend detection as well as consistency with observations (Knutson et al. 2013). None of the methods is based on seeking scenario–observation agreement in long-term trends, although this statistical property is strongly linked with short-term trend distributions. Such an agreement could be imposed on the past segment of a scenario simply by borrowing the observational regression. However, extrapolating for the future segment is unlikely to give credible results.

a. Datasets

1) NUMERICAL SIMULATIONS

Monthly data from CMIP5 are used (Taylor et al. 2012). The CMIP5 dataset has been chosen because it is recent and well explored by the climate community (Stocker et al. 2013) and reflects a variety of modeling efforts. Nonetheless, this ensemble’s representation of the historical climate has limitations (Sheffield et al. 2013), and models are not completely independent (Räisänen 2007; Masson and Knutti 2011; Stephenson et al. 2012; Knutti et al. 2013). We use 60 simulations from 15 Earth system models (ESMs) and four representative concentration pathways (RCPs). Only ESMs presenting at least one member per RCP at the time of analysis have been selected; these are listed in Table 1. The principle of “model democracy” is applied to the selected ESMs, with only the member with identification code r1i1p1 kept for each ESM and RCP (most models had only one member at the time the analysis was conducted). However, in practice, some models may have more weight than others in this study, since different models from the same center may, in fact, be closely related (GFDL and MIROC each contribute three models in this study).

2) OBSERVATIONAL DATA

The product used for scenario design is a Natural Resources Canada (NRCan) dataset based on daily observations from Environment Canada’s station network interpolated to a regular 10 km $\times$ 10 km grid over Canada (Hutchinson et al. 2009; Hopkinson et al. 2011). Version 2 is used. It contains the daily minimal ($T_{\text{min}}$)
and maximal ($T_{\text{max}}$) temperatures, obtained by applying the Australian National University Splines (ANUSPLIN) package interpolation procedure to station observations. This procedure accounts for terrain elevation but does not take account of any other physical principle. Thus, biases are potentially large where the stations network is sparse (e.g., in the Arctic). Although this dataset does not strictly correspond to observations, this term will be used herein for the sake of brevity.

Daily mean temperatures ($T_{\text{mean}}$) are obtained by the approximation $T_{\text{mean}} = (T_{\text{min}} + T_{\text{max}})/2$. This approximation was notably used by Jones et al. (1999) and has known limitations (e.g., Zeng and Wang 2012).

### b. Study domain and period

Figure 1 presents the Canadian toponyms referred to in this study. Scenarios are built for 5146 locations over the country, a number obtained by selecting each 50th point of the NRCan grid. Thus, scenarios are produced on an irregular grid, with a sampling distance of the order of $\sim 44$ km. This relatively short distance should not be interpreted as a magnification of ESMs’ resolutions ($\sim 100–300$ km); scenarios are calculated separately at each location, but neighboring locations are climatically not independent. The sampling distance stems from a compromise between limiting computing time and being able to visualize interesting geographical patterns. Each value in a time series corresponds to a single year and may represent the whole year (ANN) or a single season [December–January (DJF), March–May (MAM), June–July (JJA), and September–November (SON)]. For winter values of a given year, December data from the previous year are used, so the average value actually represents three consecutive months.

Statistical calculations of trends are performed for the 2006–35 segment of the scenarios created. During that period, the mean state of the scenarios ensemble follows a fairly linear trend (not shown), with no obvious RCP-based segregation. Figure 2 shows the linear trends for the grid point nearest to Quebec City (46.71°N, 71.21°E) during the periods 2006–35 and 2036–65, for each combination ESM–RCP (annual time series are used, which are obtained by interpolating the simulations at the NRCan grid point closest to the targeted location). The uncertainty in the trend appears to be much more dependent on ESM choice and initial conditions than on RCP choice over 2006–35. The rates of warming generally increase from 2006–35 to 2036–65 in the case of RCP8.5, are about constant for RCP6.0, and decrease for RCP4.5, whereas temperatures for RCP2.6 stabilize at warmer values than currently observed.

As a matter of indication, the $p$ value of the two-sample Kolmogorov–Smirnov test between RCP2.6 and RCP8.5 linear trend distributions is $\sim 0.14$ for 2006–35,
and <0.01 for 2036–65. Considering a 5% significance level, this means the null hypothesis that the two samples are drawn from the same parent distribution is rejected for the 2036–65 period only. Applying the Kolmogorov–Smirnov test on all 5146 locations results in the rejection of the null hypothesis 77 and 5146 times for 2006–35 and 2036–65, respectively. This supports the assumption that socioeconomic scenarios associated with RCPs do not have a major impact on trends during the study period (i.e., the warming is, in large part, already committed). Hence the dataset can reasonably be considered to include four members per ESM, and trend statistics on the scenarios may be performed without regard for the RCP. This approximation of a small forcing-related uncertainty might not hold when another ensemble of simulations is considered. Time series for the different RCPs during 2006–35 are also practically indistinguishable in terms of two other statistical properties: namely, detrended variance and lag-1 autocorrelation values (not shown).

c. Scenario design

1) METHOD 1: INTERPOLATED MODEL OUTPUTS

The first method consists of interpolating each simulation at the 5146 study locations. No statistical adjustment is performed. Spatial interpolation has been used as a form of rudimentary downscaling by Ahmed et al. (2013; see also references therein).

2) METHOD 2: STATISTICAL ADJUSTMENT OF INTERANNUAL VARIABILITY

In the second method, each simulation is adjusted according to its level of agreement with observational variability over a calibration period (1962–2010). The transient background (long-term) trends are still dictated by the simulation. Adjustment is performed independently for each simulation, season, and location. After spatial interpolation, the first step is calculating the time-varying background state of the simulation $B_{\text{SIM}}$, as a fourth-order regression (Hawkins and Sutton 2009). The $B_{\text{SIM}}$ is then subtracted from the simulated time series to

![Fig. 2. Trends in annual $T_{\text{mean}}$ time series for the 15 ESMs listed in Table 1 for periods (a) 2006–35 and (b) 2036–65, near Quebec City. Symbols corresponding to the same ESM are connected for better visual clarity.](image-url)
get the residual distribution \( \{r_{\text{SIM}}\} \). A fourth-order regression is also used to separate the observational residuals \( \{r_{\text{OBS}}\} \) from the background transient state \( B_{\text{OBS}} \). A transfer function is next determined, which ensures the mapping of \( \{r_{\text{SIM}}\} \) into \( \{r_{\text{OBS}}\} \) for elements of the calibration period, on a quantile basis. The transfer function is then applied to the 1962–2035 residuals, and a scenario is obtained by adding \( B_{\text{SIM}} \) back to the adjusted residuals.

The statistical adjustment of interannual variability is inspired by techniques applied to daily model outputs and known under different names, including quantile mapping (Thomeßl et al. 2012) and statistical bias correction (Plani et al. 2010). Here the transfer function is linked with the residuals rather than the actual values, mostly because otherwise different observed and simulated long-term trends would lead to inappropriate variability adjustment (Scherrer et al. 2005).

3) METHOD 3: STOCHASTIC INTERANNUAL VARIABILITY

In this method, simulations are used to provide the long-term background state, whereas observations are used to create year-to-year fluctuations around that state. Each simulation provides a background state \( B_{\text{SIM}} \) independently, which is determined as in method 2 (fourth-order regression). The same technique is used to obtain \( B_{\text{OBS}} \) and the corresponding residuals \( r_{\text{OBS}} \) are then used to create a first-order autoregressive or AR(1) model (Wilks 2006, section 8.3.1). In such a statistical model, each value \( X_t \) of a time series partially depends on the previous, \( X_{t-1} \), following the equation

\[
X_t = \mu + \varphi (X_{t-1} - \mu) + \epsilon_t, \tag{1}
\]

where \( \varphi \) corresponds to the lag-1 autoregressive parameter of \( \{r_{\text{OBS}}\} \), \( \mu \) is the average of \( \{r_{\text{OBS}}\} \), and \( \epsilon_t \) is a random number picked from a distribution \( \{\epsilon\} \). This distribution is assumed to be Gaussian, with \( \mu_{\{\epsilon\}} = \mu = 0 \) and \( \sigma^2_{\{\epsilon\}} = (1 - \varphi^2)\sigma^2 \), \( \sigma \) being the standard deviation of \( \{r_{\text{OBS}}\} \). Scenario residuals are then generated using the first value of \( \{r_{\text{OBS}}\} \) as the initial \( X_0 \) value and randomly picking from within \( \{\epsilon\} \). Here 50 values beyond the number needed to cover the period 1962–2035 are generated, and only the last 74 values are kept, to eliminate the influence of the arbitrary \( X_0 \) value chosen to start the series. Once the time series for the scenario residuals is obtained, it is added to \( B_{\text{SIM}} \) to obtain a \( T_{\text{mean}} \) scenario. Whether the use of a higher-order autoregressive integrated moving average (ARIMA) model would be more appropriate has not been investigated, as was done, for example, by Foster and Rahmstorf (2011) for global temperature.

Successively using each simulation of our work dataset provides an ensemble of 60 scenarios. Note that for each simulation, any number of scenarios could have been generated just by changing \( X_0 \) or the pseudorandom seed. However, using the same number of scenarios as provided by methods 1 and 2 renders the intermethod comparison easier.

4) METHODS SUMMARY

Table 2 offers an overall view of how simulations and observations are combined to produce scenarios in each method, indicating which dataset is relied upon for the background state, the sequence (alternation of minima and maxima), and the variability. The alternative method referred to in Table 2 is not used to generate future \( P_{\text{cool}} \) values and is described in section 4. Figure 3 shows what this means concretely by showing the three scenarios that stem from the FGOALS-s2 historical/ RCP6.0 simulation, for annual \( T_{\text{mean}} \) values interpolated at the grid point nearest to Yellowknife (62.13°N, −114.46°E). It is not necessary to adjust the manifest offset between simulated and observed long-term averages, because trends and \( P_{\text{cool}} \) values are invariant under a constant shift. However, to facilitate viewing, the method 3 scenario is shown with a constant shift corresponding to the difference between the means of \( B_{\text{SIM}} \) and \( B_{\text{OBS}} \) over the calibration period. The more subtle offset in variability is adjusted for in method 2. In this case, standard deviations of the 1962–2010 residuals for observations, the simulation, and the method 2 scenario are 1.15°C, 1.51°C, and 1.17°C, respectively. For method 3, the standard deviation of the residuals is 1.30°C; the number of time steps is too small to ensure convergence toward the prescribed standard deviation.

d. Short-term trend calculations

Once scenarios are created, short-term trend distributions are obtained, with the aim of estimating a probability...
for the occurrence of a negative trend \( P_{\text{cool}} \) for each case location/season/method/duration. Short-term durations considered are 5, 10, 15, 20, and 25 yr. For each case, a trend distribution is obtained by running a block of the specified duration within the 2006–35 period and calculating linear trends by least squares regression. For each case, this gives 1560 values for 5-yr durations (26 blocks \( \times \) 60 scenarios), and 1260, 960, 660, and 360 values for the other durations as they increase. It is important to note that blocks are not independent, because they overlap. As will be discussed later, this lack of independence is taken into account in assessing confidence intervals (CIs). For each distribution, some trends are negative and some are positive, and the fraction of negative values during 2006–35 is directly interpreted as \( P_{\text{cool}} \).

There are more sophisticated ways to obtain probabilities in situations like this—for example, using Bayesian approaches. However, there is no methodological consensus, and differences between results based on various methods are sometimes large (e.g., Tebaldi and Knutti 2007). It is reasonable to consider that the more frequently negative trends occur in scenarios based on ESMs, the more likely they are to occur in the real world.

### 3. Hindcast skill

#### a. Verification rank histograms

A consistent hindcast requires that the observed value’s rank varies with equiprobability among the different projected outcomes (Wilks 2006, section 7.7.2). Such an assessment has been performed for method 1 and every case season/duration during the period 1962–2010, as shown in Fig. 4. Ranks are converted into percentiles. Weights are \( 1/2N \) for each tail and \( 1/N \) for each interval between successive scenario \( P_{\text{cool}} \) values, with \( N = 60 \) being the number of scenarios. Linear interpolation is performed between ranks, and 5-percentile bins are used for the histograms (see van Oldenborgh et al. 2013 for further details). Each location contributes one \( P_{\text{cool}} \) value based on a number of trend values, which depends on the duration. Overlap in periods for calculating trends and spatial correlation limit the effective sample size in two different manners. For example, for the case ANN/10-yr duration, there are around 6 climatically independent regions across the country, and only 4 of the 40 trend values determining the local \( P_{\text{cool}} \) value are completely independent. This indicates that the histograms include some redundant information, rendering the estimation of statistical significance challenging (with respect to the disparities between observation and scenario \( P_{\text{cool}} \) distributions). Nevertheless, rank histograms are probably among the best tools for judging how observations differ from the hindcast ensemble. For each case in Fig. 4, observations (red histogram) differ from the ideal result (black line). But when individual simulations are compared with the other 59, histograms also depart from the ideal result (for each bin, the gray shading ranges from the lowest to the highest frequency.
resulting from the intersimulation verifications). The general location of the red histogram within the gray shading indicates that observations differ from simulations approximately as much as simulations differ from one another. However, for DJF and 15–25-yr durations there is an overforecasting bias (the hindcast shows perceptibly fewer low \( P_{\text{cool}} \) values than the observations).

Also, for the case SON/25 yr, observational values occur atypically often within the 80th–95th percentiles. The overall picture is the same with methods 2 and 3 (not shown).

The fairly good verification rank histograms do not highlight the fact that \( P_{\text{cool}} \) geographical patterns vary substantially between observations and individual models. Related correlation coefficients \( r_{\text{Pea}} \) (not shown) are generally small but more often positive than negative. Forcing local scenarios to adopt the observational background state improves \( P_{\text{cool}} \) correlational results substantially. This is discussed in more detail in the next section.

b. Distributions

In this section, the models’ \( P_{\text{cool}} \) distributions are compared with the corresponding observational values during the calibration period (1962–2010). This procedure’s objective is mainly to assess whether statistically adjusting the variability (with methods 2 and 3) has a beneficial impact in hindcast mode. An alternative scenario design method that consists of reproducing the observed variability around its fourth-order regression with an AR(1) model is also used. This method has not been used beyond the calibration period, because the observed regression would have to be extrapolated, and this procedure has been judged to be too simplistic. The
alternative method’s role in the current section is limited to assessing the potential for agreement between scenario and observational $P_{\text{cool}}$ values that could be reached were their background state the same. Such an agreement during the past period would not guarantee future agreement, but a disagreement would clearly indicate some limitation in the scenario design method.

In Fig. 5, winter and summer 5-yr duration $P_{\text{cool}}$ distributions for 1962–2010 across Canada are shown for observations (red histograms) as well as for the individual 15 ESMs [gray histograms; black bars connect the 4th and 12th highest frequency values for each 2.5% bin. For each case, each of the 15 scenario distributions is compared with the observational distribution through the Kolmogorov–Smirnov statistic; minimal, median, and maximal values are shown. 

Fig. 5. Histograms of $P_{\text{cool}}$ values across Canada. Each panel represents a case method/season. Each histogram is based on 5146 5-yr $P_{\text{cool}}$ values (one value per location). Red histograms are for observational values and gray histograms for scenarios from each of the 15 models. Black bars connect the 4th and 12th highest model frequencies in each 2.5% bin.
often present a $P_{cool}$ histogram with a light left tail relative to the observational one. Another noteworthy result is that there is no marked difference between methods 1 and 2, as revealed graphically and by $D_{KS}$ values. Method 3 leads to model histograms much closer to one another for cases shown in Fig. 5. One potential explanation is spatial decorrelation of the variability, which produces much richer effective sampling. The difference between results for method 3 and those for the alternative method previously described illustrates how the right background state may contribute to obtaining the right $P_{cool}$ values.

For a more complete picture, Fig. 6 presents the $D_{KS}$ values for each case season/duration/method/ESM. Each model’s $P_{cool}$ distribution is compared with the observational one. Again, each black bar gives the IQR among the 15 ESMs. The green marker corresponds to $D_{KS}$ between observations and the multimodel ensemble. Results can be summarized as follows. Adjusting variability with method 2 generally does not bring $P_{cool}$ distributions closer to the observed ones. This results partially from compensational effects, whereby occurrence numbers for inflation and compression of standard deviation are comparable (this is not the case for winter,

---

**Fig. 6.** Kolmogorov–Smirnov distance between observational and scenario distributions for each case season/duration/method. Each black bar connects the 4th and 12th highest values among the 15 ESMs (distributions from the different RCPs are concatenated). A green marker corresponds to $D_{KS}$ between observations and the multimodel ensemble.
where models produce many more cases of excessive variability than the reverse). Altered trends must also cross the zero line in order to make a difference in $P_{\text{cool}}$, so it is in some ways harder to impact $P_{\text{cool}}$ than to impact the underlying trend distribution. Generating variability with an AR(1) model (in method 3) often—but not always—leads to a narrower IQR, but this does not necessarily co-occur with a general $D_{KS}$ decrease. Thus, altering variability and/or lag-1 autocorrelation toward observed values does not systematically lead to closer $P_{\text{cool}}$ distributions, although very high $D_{KS}$ values are often diminished with method 3. The multimodel distribution is found most often within the IQR of individual models and sometimes in the best quartile. There is no case where it is found in the worst quartile.

Results for the alternative method suggest at least two things. First, using the same background state as observations often brings a more substantial improvement than using the same standard deviation. Second, with the same background states, it seems that 4 AR(1) runs per location are enough for having convergence toward some $P_{\text{cool}}$ distribution (this is not the case for an individual location). The spread associated with method 3 is thus more due to the simulations having diverse offsets in their background state than to random effects of the AR(1) procedure. One case in which the alternative method clearly results in worse $D_{KS}$ values is the DJF/10-yr case (annual results are impacted as well). For this case, scenario histograms are shifted by ~4% toward lower $P_{\text{cool}}$ values, because of a wide region around Hudson Bay where the observed $P_{\text{cool}}$ values are much larger than expected from the alternative method (not shown). In this region, standard deviation of the observed residuals has increased over time, and the AR(1) model cannot systematically reproduce this heteroscedasticity. Method 3 thus obtains good results for the wrong reason (incorrect scedasticity partially compensating for inadequate background states). Another potential inadequacy of an AR(1) model could stem from residuals not being Gaussian (Gluhovsky and Agee 2007).

The availability of metrics to assess how realistic simulated $P_{\text{cool}}$ distributions are during the recent-past period renders the recourse to unequal model weights tempting. For example, a relative weight equal to $1 - D_{KS}$ could be a simple, defensible choice. However, the principle of model democracy used here has been preserved for two reasons. First, better past skill does not guarantee better future skill, because the relative importance of the physical processes can evolve in a transient climate (Reifen and Toumi 2009). Second, relative model performances vary substantially across seasons and trend durations (but less across methods). For example, each raw simulation is found in the best quartile for at least one case season/duration, and the result for the worst quartile is similar.

In this section, it has been shown that forcing simulated standard deviations to fit the observed values does not lead to systematic benefits or shortcomings in
hindcasting $P_{\text{cool}}$ distributions. This occurs because other statistical properties, such as long-term trend and scedasticity, are also important. Taking this into account, results for 2006–35 (next section) have been computed using each of the three methods, and differences are interpreted as uncertainty related to post-processing of simulated interannual variability.

4. Estimation of 2006–35 $P_{\text{cool}}$ values

a. A specific case (Quebec City)

Figure 7 shows smoothed histograms (distributions) of the trends during 2006–35 near Quebec City, for the five durations considered. Annual values and method 2 are used for this figure, with bins of size 0.333°C decade$^{-1}$ (the smoothing procedure is to facilitate viewing and consists of connecting the bin centers). Unsurprisingly, medians are located in the positive trend zone (between 0.38°C and 0.47°C decade$^{-1}$), the shorter the duration is the wider the distribution becomes, and the fraction of the area under the curve located in the negative zone decreases as duration increases. Corresponding $P_{\text{cool}}$ values are presented in Fig. 8, along with results for each season and for the other scenario design methods. Square symbols indicate the $P_{\text{cool}}$ value (frequency of occurrence), whereas associated solid lines illustrate CIs at the 95% level (see the appendix). The gradual decrease in $P_{\text{cool}}$ as duration increases appears to be robust through seasons and methods. Values approach 50% for the 5-yr duration, but this threshold is not exceeded (however, this does happen at a few locations for the 5-yr duration and method 3, with the extremal case at $P_{\text{cool}} = 51.9\%$). The MAM and DJF seasons generally show the highest $P_{\text{cool}}$ values, whereas ANN and JJA generally show the lowest values.

It is important to mention that the method for estimating CIs only accounts for stochastic uncertainty, through the assumption that each sample is drawn from a Gaussian
distribution. Other issues, including model interdependence and RCP representativeness, are not accounted for by the CIs. Dependence on the postprocessing method is, however, explicitly shown in the three panels. The variety of error sources associated with estimating probabilities strongly suggests not interpreting \( P_{\text{cool}} \) values too precisely. Hence, there must be more focus on geographical patterns and comparisons between seasons, methods, and durations than on the \( P_{\text{cool}} \) values themselves. CIs will not be presented for the remainder of the paper, for the sake of brevity.

\textbf{b. Results on the Canadian scale}

Probabilities of negative trends have been calculated for the 5146 grid points covering Canada. Figures 9, 10, and 11 show results obtained per season for the 20-yr duration using methods 1, 2 and 3, respectively. Patterns are relatively similar between methods, although maps appear noticeably noisier for method 3, because of the fact that variability is spatially uncorrelated. Roughly speaking, the regions with the highest \( P_{\text{cool}} \) values are parts of the Prairies (as defined in Fig. 1), British Columbia, southern Yukon, and Newfoundland–Labrador for DJF, Newfoundland–Labrador and the Eastern Baffin Island area for MAM, land surrounding the Beaufort Sea and a thin segment along the Labrador coast for JJA, and parts of the Prairies for SON. On the other hand, the regions with the lowest \( P_{\text{cool}} \) values correspond to parts of Nunavut for DJF and MAM, to southern Quebec, New Brunswick, and parts of the Canadian Cordillera for JJA, and to southeastern Canada and parts of Nunavut and southern Ontario for SON.

Increasing (decreasing) the duration generally leads to the same geographical patterns, with an overall decrease (increase) in \( P_{\text{cool}} \) (not shown). This overall variation linked with duration is illustrated in Table 3, which presents the \( P_{\text{cool}} \) interdecile range [IDR (10th–90th percentiles)] of the distribution of the 5146 values over Canada, for each case duration/season/method. As duration increases (for fixed season and method), the IDR moves toward lower probability values, and generally widens. With regards to the method’s impact (keeping duration and season fixed), method 3 often results in IDRs with the lowest probability values. Finally, seasonal influence could roughly be characterized by the following ranking of \( P_{\text{cool}} \) for fixed duration and method: MAM, DJF, JJA, SON, and ANN (from highest to lowest values, based on 10th and 90th percentile values). Spring and winter are associated with
higher probabilities on the Canadian scale, but this is not necessarily the case for every location.

Results provided here consist of the statistical aggregate of a variety of climate trajectories, which, moreover, stem from models with different (though related) physics formulations. A physical interpretation would necessitate an investigation of physical processes underlying the trajectories of individual simulations, as performed, for example, by Deser et al. (2014). Such an analysis is beyond the scope of the present paper, but how variability and long-term linear trend patterns correlate with $P_{\text{cool}}$ patterns has been investigated. Results using Spearman rank correlation coefficients ($r_{\text{Spe}}$) are shown in Table 4 for the 5-yr duration. For each case season/method/location, 30-yr linear regressions are performed on each of the 60 scenarios to determine individual variability and trend values, which are next averaged. Results verify a posteriori that variability is generally linked with $P_{\text{cool}}$, although the low number of climatically independent regions implies a low level of statistical significance. The long-term trend appears more important, and the variability-to-trend ratio is an even better indicator (absolute values of $r_{\text{Spe}}$ are the same when the inverse signal-to-noise ratio is used). Results are qualitatively the same for longer durations but have not been shown since it is trivial that, for example, 25-yr $P_{\text{cool}}$ values are correlated with 30-yr linear trends. These results suggest that, for example, the low winter (high summer) $P_{\text{cool}}$ values for the western Arctic Archipelago relative to other regions (Figs. 9–11) may be explained in terms of relatively low (high) ratios of the variability over the background trend.

5. Summary with discussion

This paper aims to produce probabilistic scenarios for short-term cooling trend occurrence over Canada during 2006–35 and to evaluate the impact of climate variability postprocessing on results. The frequency of negative temperature trends in an ensemble of 60 climate scenarios has been interpreted as the probability of cooling trend ($P_{\text{cool}}$), using three different postprocessing methods: direct usage of locally interpolated simulations (method 1), calibration of variability based on a procedure akin to bias correction (method 2), and use of an autoregressive model (method 3).

It has been verified in hindcast mode (during 1962–2010) that $P_{\text{cool}}$ distributions across Canada are relatively reliable. Indeed, observational distributions do not differ from scenario distributions more than these...
differ from one another, except for winter and 15–25-yr durations (atypical overforecasting bias) and for autumn and 25-yr duration (underforecasting). Hindcast results also show that postprocessing of the variability has a slight impact on $P_{\text{cool}}$ distributions but does not improve reliability much. Moreover, it appears that short-term $P_{\text{cool}}$ reliability is more limited by inadequacies in model long-term trends than in model variability. Results for 2006–35 suggest that the most influential indicator may be the signal-to-noise ratio, because short-term $P_{\text{cool}}$ across Canada is more closely correlated to the ratio of the background trend over variability than to the background trend or to variability alone.

Calculations for 2006–35 unsurprisingly lead to higher $P_{\text{cool}}$ values for shorter durations. For example, probabilities of cooling in mean annual values across Canada decrease from ~40%–46% for 5-yr durations to ~2%–18% for 25-yr durations (interdecile ranges). The case of Quebec City highlights that stochastic uncertainty in $P_{\text{cool}}$ is considerable and may be larger than that associated with the interannual variability postprocessing method. However, it must be mentioned that a larger ensemble would have led to less stochastic uncertainty and that the assigned number of effective degrees of freedom is somewhat subjective, in part because models are not fully independent.

It must be emphasized that besides the stochastic issue, there is also epistemic uncertainty. Indeed, the CMIP5 simulations used are based on four different representative concentration pathways (RCPs), which...
do not cover all socioeconomic possibilities. For example, the deployment of a geoengineering scheme involving sulfate aerosols (Keith 2013) or a nuclear conflict (Ozdoğan et al. 2013) are two conceivable events for which regional climate impacts could be outside of the range covered by RCP-based simulations. Volcanism and solar activity, not accounted for in RCPs because they are natural forcings, could also impact $P_{\text{cool}}$ calculations, were they more predictable. One interesting approach would be to use occurrence probabilities for significant volcanic eruptions (Hyde and Crowley 2000) and for the onset of a grand solar minimum (Lockwood 2010).

Despite the limitations just mentioned, it seems clear that, during the coming decades, cooling trends of durations up to 25 yr (and possibly more) do represent plausible outcomes at any location in Canada. This in no way disproves global warming, because these cooling trends occur in simulations whose background state is warming. It also appears that, as a counterpoint, short-term warming trends could occur that are much stronger than the expected range of long-term trends.

Acknowledgments. We acknowledge the World Climate Research Programme’s Working Group on Coupled Modelling, which is responsible for CMIP, and we thank the climate modelling groups (listed in Table 1 of this paper) for producing and making their model output available. The U.S. Department of Energy’s Program for Climate Model Diagnosis and Intercomparison provides coordinating support and led the development of software infrastructure for CMIP in partnership with the Global Organization for Earth System Science Portals. We also thank Dan McKenney and his team at the Canadian Forest Service of Natural Resources Canada for providing the observational product used here. Finally, we acknowledge the useful work of three anonymous reviewers.

### Table 4. Spearman rank correlation coefficient between 5-yr $P_{\text{cool}}$ and other statistical properties of scenarios over 5146 grid points of Canadian inland territory during 2006–35. Variability (standard deviation of linearly detrended time series) and linear trends are calculated for each scenario and then averaged over the 60 scenarios at each grid point before variability/trend ratios are calculated and correlated with $P_{\text{cool}}$.

<table>
<thead>
<tr>
<th></th>
<th>Variability</th>
<th>1/trend</th>
<th>Variability/trend</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Annual</td>
<td>0.40</td>
<td>0.36</td>
<td>0.39</td>
</tr>
<tr>
<td>DJF</td>
<td>0.23</td>
<td>0.46</td>
<td>0.63</td>
</tr>
<tr>
<td>MAM</td>
<td>0.29</td>
<td>0.49</td>
<td>0.53</td>
</tr>
<tr>
<td>JJA</td>
<td>0.14</td>
<td>0.42</td>
<td>0.56</td>
</tr>
<tr>
<td>SON</td>
<td>0.00</td>
<td>0.06</td>
<td>0.07</td>
</tr>
</tbody>
</table>

### APPENDIX

#### Confidence Intervals for $P_{\text{cool}}$

Uncertainty in $P_{\text{cool}}$ is determined by creating a model of the population of trends for each case based on considerations found in Schaeffer and Mcclave (1990, section 7.3). To explain the method used, let us consider the sample of trends for the 10-yr duration shown in Fig. 4. This distribution is based on 1260 values, corresponding to 60 independent simulations and 21 blocks falling within 2006–35 (the first 10-yr block would be 2006–15), and has a mean $m$ and a variance $s^2$. The sample is assumed to be drawn from a population with a Gaussian distribution with a mean $\mu$ and variance $\sigma^2$, so that $P_{\text{cool}}$ of the population is given by

$$P_{\text{cool}} = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi} \sigma} e^{-(x-\mu)^2/2\sigma^2} \, dx. \quad (A1)$$

The parameters $\mu$ and $\sigma$ cannot be determined unequivocally from $m$ and $s$, and confidence intervals must be set, at the 95% level in this case. The CI for $\sigma^2$ is given by

$$[(n-1)s^2/\chi_{0.025}^2, (n-1)s^2/\chi_{0.975}^2]. \quad (A2)$$

where $n-1$ stands for the number of degrees of freedom and $\chi_{0.025}^2$ and $\chi_{0.975}^2$ represent the corresponding critical values for a chi-square distribution, appropriate for the variance of a Gaussian distribution. A conservative minimum of 60 degrees of freedom is considered, because blocks overlap and to account for possible autocorrelation. The CI for $\mu$ is normally given by

$$[m-t_{0.025} s/\sqrt{n}, m+t_{0.975} s/\sqrt{n}]. \quad (A3)$$

where $t_{0.025}$ and $t_{0.975}$ represent the critical values for a Student’s $t$ distribution, appropriate for the mean of a Gaussian distribution. As $n > 30$, the $t$ distribution approaches the Gaussian distribution, and the critical values $t_{0.025} = t_{0.975} = 1.96$ are used. In the present case, $s$ is conservatively replaced by the upper bound of the CI for $\sigma$.

Once the CIs for $\mu$ and $\sigma$ are determined, the four combinations of the CI bounds are successively introduced in Eq. (A1), and the two extremal values among these four are used as the CI for $P_{\text{cool}}$. Consistency is ensured by checking that $P_{\text{cool}}$, calculated as the proportion of negative values in the sample, falls within the corresponding CI. For the case referred to above, the sample has $m = 0.41 ^\circ C$ and $s = 0.84 ^\circ C$, which gives a CI for $P_{\text{cool}}$ of [17.2%, 43.9%], whereas the proportion of negative values in the sample is 30.4%.


