Clim Dyn
DOI 10.1007/s00382-012-1330-3

Higher precision estimates of regional polar warming by ensemble regression of climate model projections

Thomas J. Bracegirdle · David B. Stephenson

Received: 30 August 2011 / Accepted: 29 February 2012

Abstract This study presents projections of twenty-first century wintertime surface temperature changes over the high-latitude regions based on the third Coupled Model Inter-comparison Project (CMIP3) multi-model ensemble. The state-dependence of the climate change response on the present day mean state is captured using a simple yet robust ensemble linear regression model. The ensemble regression approach gives different and more precise estimated mean responses compared to the ensemble mean approach. Over the Arctic in January, ensemble regression gives less warming than the ensemble mean along the boundary between sea ice and open ocean (sea ice edge). Most notably, the results show 3 °C less warming over the Barents Sea (~7 °C compared to ~10 °C). In addition, the ensemble regression method gives projections that are 30 % more precise over the Sea of Okhostk, Bering Sea and Labrador Sea. For the Antarctic in winter (July) the ensemble regression method gives 2 °C more warming over the Southern Ocean close to the Greenwich Meridian (~7 °C compared to ~5 °C). Projection uncertainty was almost half that of the ensemble mean uncertainty over the Southern Ocean between 30° W to 90° E and 30 % less over the northern Antarctic Peninsula. The ensemble regression model avoids the need for explicit ad hoc weighting of models and exploits the whole ensemble to objectively identify overly influential outlier models. Bootstrap resampling shows that maximum precision over the Southern Ocean can be obtained with ensembles having as few as only six climate models.

Keywords CMIP3 · CMIP5 · Climate model · Arctic · Antarctic · Regional climate · Weighting · Observational constraint · Southern Ocean · Sea ice edge · Polar climate

1 Introduction

The uncertainty in climate model projections is particularly large near the boundary between sea ice and open ocean (referred to hereinafter as the sea ice edge) (e.g. Tebaldi et al. 2005; Greene et al. 2006; Christensen et al. 2007). Precise climate change response estimates on the local (grid-box) scale are important for impact studies of the socio-economic consequences of future change. In particular, changes in the polar regions can have a global impact through changes in sea level and ocean circulation. Precision is a well-known concept in science and engineering that is distinct from accuracy. Accuracy is how close the estimated response is to the future observations whereas precision is the sampling uncertainty in our estimate. High precision is a necessary but not sufficient condition for reliable accuracy, since the mean estimate might have a small sampling uncertainty but might not be centred around the true future value. In previous studies, methods for improving the precision of estimated future climate change over the polar regions have often involved ad-hoc weighting to remove climate models with large local biases in their simulation of observed climate (e.g. Overland and Wang 2007; Stroeve et al. 2007; Bracegirdle et al. 2008;

T. J. Bracegirdle (✉)
British Antarctic Survey, High Cross, Madingley Road, Cambridge CB3 0ET, UK
e-mail: tjbra@bas.ac.uk

D. B. Stephenson
Mathematics Research Institute, University of Exeter, Exeter, UK

D. B. Stephenson
NCAS-Climate, Reading, UK

Published online: 17 March 2012

Walsh et al. 2008; Zhang 2010). However, there are inherent difficulties in estimating and justifying the weights applied to different climate models.

The simplest form of model averaging is to weight all climate change responses equally by calculating the ensemble mean response. However, this fails to account for poorly performing models and so it is common practice to consider ensemble means of only a subset of models (i.e. assign zero weight to some models). For example, Arctic sea ice declines have been estimated by discarding climate models with large biases in sea ice extent (Overland and Wang 2007; Stroeve et al. 2007; Zhang 2010). However, the decision over what threshold to use for discarding a climate model from the original ensemble can be difficult to justify and risks leaving out models that will still contribute valuable information. More complex methods for assigning weights to climate models have been proposed over the last decade, based largely on weighting according to how well the model reproduces the observed mean climate (Giorgi and Mearns 2002; Murphy et al. 2004; Connolley and Bracegirdle 2007). An important caveat with these approaches is that they do not consider the extent to which biases in simulated present day mean climate may be related to climate change responses (Raisanen et al. 2010). Recently there has been a greater consideration of the importance of this over the polar regions (Whetton et al. 2007; Raisanen et al. 2010; Abe et al. 2011).

Whetton et al. (2007) found significant similarity between patterns of regional present day climate and patterns of regional future change. However, Watterson and Whetton (2011) found that weights based on the '*M*' similarity statistic used by Whetton et al. (2007) have little impact on multi-model PDF spread. Abe et al. (2011) used an alternative approach in which singular value decomposition (SVD) was applied to extract modes relating inter-model variability in present day climate to variability in future change. When compared to projected climate change based on an equal-weight ensemble mean, their method shows increased precision over the Arctic, but decreased precision at lower latitudes. These mixed results may be a consequence of defining the model weights based on non-local global-scale patterns. Raisanen et al. (2010) took a more local approach by deriving weights at each model grid point from present-day-future relationships of key variables. They showed increased precision near the sea ice edge in winter and negligible differences elsewhere. However, estimation of weights from such relationships remains rather difficult due in part to the small size of most climate model ensembles (Knutti et al. 2010; Weigel et al. 2010).

Boe et al. (2009) took the less complicated approach of directly using the estimated mean response from linear regression applied to the relationship between simulated recent total Arctic sea ice extent trends and projected extent in the early-to-mid twenty-first century. Estimates derived from linear regression onto such relationships have been presented based on other large-scale parameters, such as Northern Hemisphere snow-albedo feedback (Hall and Qu 2006) and Arctic-wide warming (Mahlstein and Knutti 2011). A key issue associated with the polar regions is the large internal variability of the climate (Wang et al. 2007), which potentially introduces a large amount of additional uncertainty into estimated mean climate change responses. In addition, the large-scale parameters used in the above studies are not necessarily representative of specific locations that are important for impact studies.

In this study, we introduce a simpler robust statistical framework that addresses these issues and apply it to the Coupled Model Inter-comparison Phase 3 (CMIP3) dataset (Table 1). There are three important elements in this framework. Firstly, local present day mean climate is used as a predictor for the future climate change response, based on linear regression of inter-model relationships at each model grid point. Secondly, a procedure for identifying influential climate models having large leverage in the regression is introduced. Thirdly, a procedure is introduced to determine the point at which errors stop decreasing with increasing ensemble size (or whether a larger ensemble is required). Together these three elements provide a new framework for producing more precise climate change projections at the local (grid box) scale. We refer to this statistical model-based approach as ensemble regression (ER). For locations where the multi-model climate change response is uncorrelated with present day climate, the ER approach effectively reverts to an ensemble mean (EM) approach.

Section 2 of this paper describes the statistical methodology and data sources. Arctic and Antarctic projections of near-surface winter temperature change over the twenty-first century estimated using ER are then shown in Sect. 3 and compared to projections estimated using the EM approach (see Fig. 1 for the Arctic and Antarctic locations referred to in this paper). Section 3 includes results from a quadratic regression model (in Sect. 3.4), which is shown to give less reliable results than those based on linear regression. Section 4 concludes with a summary and discussion of possible future extensions.

2 Statistical methodology and data

2.1 Ensemble mean

The simplest approach for inferring the observable climate change response from multi-model ensembles of climate projections is to take the arithmetic mean of all the climate model responses. In other words, the estimate of the

Table 1 IPCC CMIP3 models used in this study

Model ID	Model name	20c3m runs	Sresa1b runs	Institute
1	BCCR BCM2.0	1	1	Bjerknes Centre for Climate Research
2	CCSM3	1,2,3,4,5,6,7,8	1,2,3,4,5,6,7	National Center for Atmospheric Research
3	CGCM3.1(T47)	1,2,3,4,5	1,2,3,4,5	Canadian Centre for Climate Modelling and Analysis
4	CGCM3.1(T63)	1	1	Canadian Centre for Climate Modelling and Analysis
5	CNRM-CM3	1	1	Centre National de Recherches Meteorologiques
6	CSIRO-Mk3.0	1,2,3	1	Commonwealth Scientific and Industrial Research Organisation (CSIRO) Atmospheric Research
7	CSIRO-Mk3.5	1,2,3	1	Commonwealth Scientific and Industrial Research Organisation (CSIRO) Atmospheric Research
8	ECHAM5/MPI-OM	1,2,3,4	1,2,3,4	Max Planck Institute for Meteorology
9	ECHO-G	1,2,3,4,5	1,2,3	Meteorological Institute of the University of Bonn, Meteorological Research Institute of KMA, and Model and Data group
10	FGOALS-g1.0	1,2[a],3	1,2,3	LASG/Institute of Atmospheric Physics
11	GFDL-CM2.0	1	1,2,3	Geophysical Fluid Dynamics Laboratory
12	GFDL-CM2.1	1	1,2	Geophysical Fluid Dynamics Laboratory
13	GISS-AOM	1,2	1,2	NASA/Goddard Institute for Space Studies
14	GISS-EH	1,2,3,4,5	1,2,3	NASA/Goddard Institute for Space Studies
15	GISS-ER	1,2,3,4,5,6,7,8,9	1,2,3,4,5	NASA/Goddard Institute for Space Studies
16	INGV-SXG	1	1	Instituto Nazionale di Geofisica e Vulcanologia
17	INM-CM3.0	1	1	Institute for Numerical Mathematics
18	IPSL-CM4	1,2	1	Institut Pierre Simon Laplace
19	MIROC3.2(hires)	1	1	Center for Climate System Research
20	MIROC3.2(medres)	1,2,3	1,2,3[a]	Center for Climate System Research
21	MRI-CGCM2.3.2	1,2,3,4,5	1,2,3,4,5	Meteorological Research Institute
22	PCM	1,2,3,4	1,2,3,4	National Center for Atmospheric Research
23	UKMO-HadCM3	1	1	Hadley Centre for Climate Prediction and Research/UK Met Office
24	UKMO-HadGEM1	1	1	Hadley Centre for Climate Prediction and Research/UK Met Office

[a] These runs were found to include erroneous values of near-surface temperature and were therefore omitted from the ensemble averages

observable response $\hat{y}_0 = \hat{x}'_0 - \hat{x}_0$ provided by the ensemble mean of the n climate model projections is:

$$\hat{y}_0 = \hat{x}'_0 - \hat{x}_0 = \frac{1}{n}\sum_{i=1}^{n}(x'_i - x_i) \quad (1)$$

where x_i is the mean present day climate in the ith model, x_0 is the present day observed climate (e.g. the 1970–1999 mean), x' denotes future climate (e.g. the 2070–2099 mean), and the hat, $\hat{}$, symbol denotes an estimate or prediction[1] of a random variable. Throughout this paper grid point near-surface temperature values are used for x_i and x_0. Where spatial means of $\hat{y}_0$ are shown, they are calculated as area-weighted averages of values at individual grid points within the area of interest. Although very easy to implement and explain, the equal weighting of model responses is not robust to overly influential outlier models. The use of equal weights is justified if one can assume that all the responses can be well described by the following statistical model:

$$y_i = \mu + \varepsilon_i \quad \text{for} \quad i = 0, 1, \ldots, n \quad (2)$$

where ε_i is an identically independently distributed random variable with zero expectation (i.e. noise). The ε_i stochastically represents model uncertainty in the response. The ensemble mean can easily be shown to provide an unbiased estimate of the mean response μ parameter in this model. For normally distributed noise, the ensemble mean is also the maximum likelihood estimate of μ for this statistical model. But are the assumptions of this simple statistical model justified?

2.2 Ensemble regression

Projected regional climate responses can depend on the model's basic present day state. For example, the surface

[1] In this paper the word prediction is used in the statistical sense to signify the expectation of the response variable for a given explanatory variable obtained using a regression model. It does not necessarily refer to future forecasts.

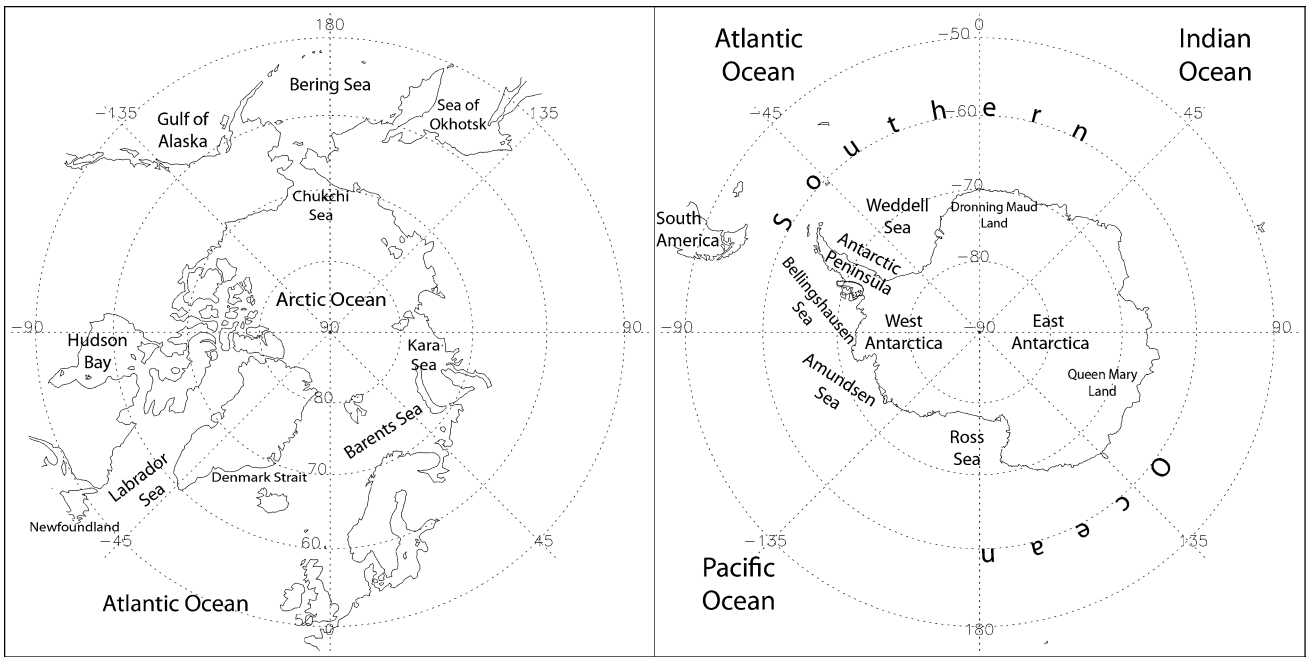

Fig. 1 Maps of the Arctic and Antarctic that show key place names referred in the main text

warming response over polar regions strongly depends on the presence of sea ice in present day model simulations (Raisanen 2007; Knutti et al. 2010). The state-dependency of the response is not represented in the ensemble mean statistical model in Eq. (1). Such unrepresented variation in the mean response leads to dependency in the ε_i, which then violates the model assumptions. The ensemble mean model is therefore inconsistent (i.e. mis-specified) if there is state-dependency in the climate response.A more flexible approach is to relate the responses to the basic present day state using a linear regression model such as

$$y_i = \mu + \beta x_i + \varepsilon_i \quad \text{for} \quad i = 0, 1, \ldots, n \tag{3}$$

where ε_i is an identically independently distributed random variable with zero expectation (i.e. noise). The Ensemble Mean (EM) model (referred to as MMM in Raisanen et al. 2010) is a special case of the ensemble regression (ER) model when $\beta = 0$. The model parameters μ, β, and σ_ε^2 (the variance of ε) can easily be estimated using ordinary least squares (Draper and Smith 1998).

Examples of the use of least squares linear regression as applied in ER are presented in Fig. 2. Scatter plots of winter near-surface temperature climate change response y_i versus present day mean x_i are shown at selected locations along the Greenwich Meridian (based on the CMIP3 climate models see Sect. 2.4 and Table 1). Each star represents the average of all available ensemble members from a given model. Significant linear association (state dependency) is clearly visible in scatter plots of locations near the sea ice edge (Fig. 2b, e). At 60° S, 0° E (Fig. 2e) the relationship is very strong ($r^2 = 0.77$) showing that in most, but not all, cases models with a colder present-mean give a larger future warming. At 75° N, 0° E (Fig. 2b) the relationship is also significant, but with a smaller slope. In addition, model 10 is a clear outlier with present day mean T_s approximately 20 °C lower than any other model. The issue of influential outliers is assessed across both polar regions in Sect. 2.3. At mid-latitudes the r^2 values are small and the regression slopes are not significantly different from zero (Fig. 2c, f). Poleward of the sea ice edges the picture is less clear and more difficult to interpret physically. At the North Pole (Fig. 2a), although small, it is notable that the slope is of an opposite sign to locations near the sea ice edge. At the South Pole (Fig. 2d) the slope is significant, but much smaller than at 60° S, 0° E. The general picture of strong correlations in T_s at high latitudes is broadly representative of other longitudes beyond the examples shown here (e.g. see spatial maps of correlations shown in Fig. 3 of Raisanen (2007) and also Fig. 7 of Knutti et al. (2010)). An implicit implication of the regression relationships shown in Fig. 2 is that at locations with a strong state-dependency, models that are closer to present day observed basic state are more likely to be closer to future observed climate change (assuming the same emissions scenario both in models and in observations).

The main mechanism for the strong state-dependency near the sea ice edge in winter is likely to be the strong local correspondence between near-surface temperature

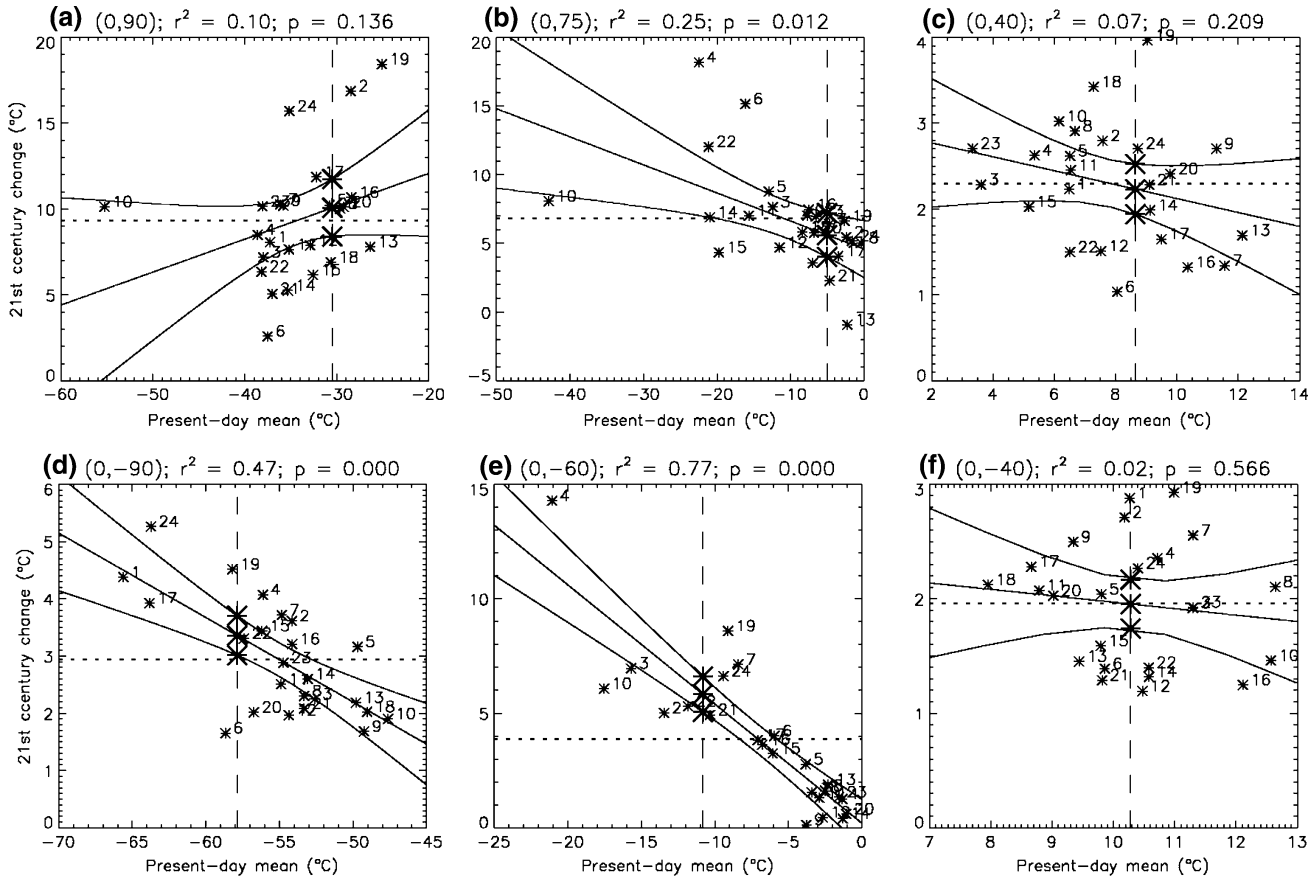

Fig. 2 Scatter plots of twenty-first century projected changes versus present day means in wintertime near-surface temperatures at locations along the Greenwich Meridian: *the top panels* **a**–**c** show Northern Hemisphere locations in January, *bottom panels* **d**–**f** show Southern Hemisphere locations in July. *Left panels* **a**, **d** show temperatures at the poles, *middle* **b**, **e** at the sea ice edges, and *right* **c**, **f** at mid-latitudes. The longitude and latitude are indicated in *brackets* at the *top left* of each plot along with r^2 and p values. The p values were calculated using a 2-tailed t test and show whether the null hypothesis (that the slope is zero) can be rejected at a given significance level. Each *small asterisk* represents one CMIP3 climate model, which are annotated by the numbers used as identifiers in Table 1. The *straight line* fits are from linear regression and the *solid curves* show the 95 % confidence intervals. The *vertical dashed lines* show the present day observations (ERA-40 data) with *large asterisks* showing the associated mean response and confidence interval from linear regression. The *horizontal dotted line* shows the simple equal-weight multi-model average of twenty-first century change

and sea ice fraction. The retreat of sea ice over the tewnty-first century leads to dramatic regional warming as the winter atmosphere is exposed to the relatively warm open ocean. Models with excessive local sea ice will therefore show more warming as the ice retreats in the future. This effect was demonstrated by Holland and Bitz (2003), who showed that CMIP2 models with a larger present day Arctic sea ice extent give a more equatorward maximum in T_s warming, since the transition from sea ice to open ocean occurs further south in those models. Related to this, Abe et al. (2011) found that, in the CMIP3 models, future regional Arctic warming is significantly related to local changes in sea ice concentration. Another possible mechanism for the state-dependency is a relationship between ocean heat transport and Arctic warming that has been suggested by Mahlstein and Knutti (2011). However, this is probably more apparent as a non-local effect and would not necessarily contribute significantly to local state-dependency in T_s. A full assessment of potentially important explanatory variables, local and non-local, is beyond the scope of this paper, but is a priority for future work.

Projections based on the application of ER to state-dependency relationships involve several important simplifying assumptions:

- As in Whetton et al. (2007) and Raisanen et al. (2010), it is assumed that there is a systematic linear relationship (with a unique slope parameter) between present day mean and future change shared universally across the different climate models and observations. Therefore, results of both the ER and EM approaches are susceptible to the effects of potentially important missing processes such as black carbon deposition (Shindell and Faluvegi 2009);

- Bias and its relationship to the basic state remains stationary into the future. Unlike the EM approach, some future changes in bias can be accommodated by the ER approach if they are related to the present day basic state. These issues are discussed in detail in Ho et al. (2012);
- The errors about the line of best fit are assumed to be identically and independently distributed;
- Climate model results and observations are considered to be interchangeable. The effects due to observational measurement error and model bias are considered to be negligible compared to other sources of uncertainty.

These assumptions are less restrictive than those of the EM approach and can be tested in various ways as will be shown later in this paper. The inclusion in the model of a systematic dependence on the basic state helps to yield residual errors that are identically and independently distributed.

Our ER approach differs from previous regression approaches in several important ways. Raisanen et al. (2010) used weights based on linear regression fits to pairwise differences between models and observations. From Eq. (2), it can be seen that $y_i - y_j = \mu + \beta(x_i - x_j) + \varepsilon_i - \varepsilon_j$ and so their linear model is equivalent to our simpler formulation. Unlike our approach where the weights emerge naturally from the mean regression-prediction of the statistical model, the weights in Raisanen et al. (2010) involve additional subjective parameters. Boe et al. (2009) used a simpler approach to ours that used a regression of future sea ice extent on recent 1979–2007 sea ice trends. Unlike our approach, their linear model predicted the expected value of the future variable rather than expected changes between future and present day values. Use of changes is advantageous since it can help correct for some of the individual model biases.

2.3 Estimated mean response

For each grid point, projections based on the ensemble regression model are given by the following estimate of the expected observable mean climate change response:

$$\hat{y}_0 = \hat{\mu} + \hat{\beta}x_0 = \bar{y} + \hat{\beta}(x_0 - \bar{x}). \quad (4)$$

This expression can easily be rewritten as a weighted sum of the model responses $\hat{y}_0 = \sum_{i=1}^{n} w_i y_i$ with grid point weights in each model given by

$$w_i = n^{-1}\left(1 + (x_i - \bar{x})(x_0 - \bar{x})/s_x^2\right) \quad (5)$$

where s_x^2 is the sample variance of the present day temperatures from the climate models and $\bar{x} = n^{-1}\sum_{i=1}^{n} x_i$. The weights w_i will generally differ from $1/n$ but the mean response estimate $\hat{y}_0$ will always be equal to $\bar{y}$ when there is no correlation between x_i and y_i (note that negative values of w_i are possible). Figure 3 shows that many of the CMIP3 models give spatially averaged grid point w_i values that differ significantly from $1/n$ both globally and over the Arctic and Antarctic (area averages of grid point weights are shown since separate regressions are made at each grid point and thus 23 global maps would be required for all weights to be displayed). Over middle and lower latitudes where x_i and y_i are generally uncorrelated (Raisanen 2007; Knutti et al. 2010), $\hat{y}_0$ largely reverts to the equal-weight mean response $\bar{y}$ (e.g. Fig. 2e, f). The weights themselves do not involve the term $x_i - x_0$ and so do not depend explicitly on the distance between simulated and observed present day climate. In other words, models closer to the observations in present day do not necessarily get higher weights. It can easily be shown that if the observations coincide with a present day model value i.e. $x_0 = x_j$, then the weights are the jth column (or row) in the hat matrix (see "Appendix"). Unlike weighting based on ad hoc metrics, these weights have emerged naturally from predictions of a statistical model, which is based on assumptions that can be rigorously tested.

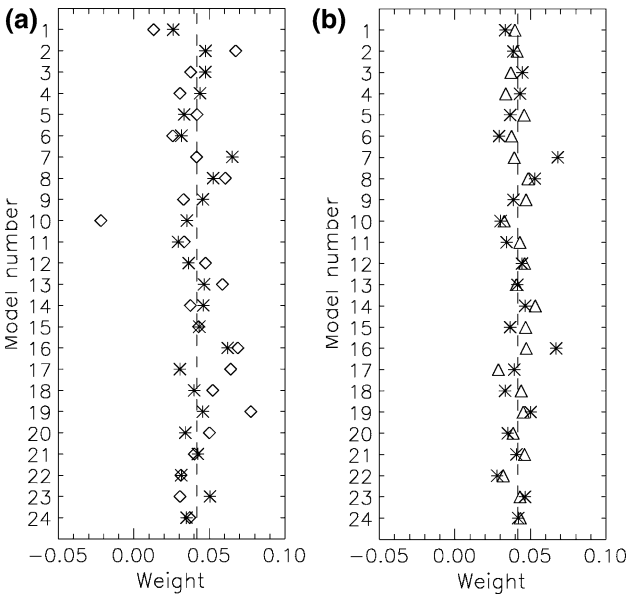

Fig. 3 Area-averaged weights for each of the CMIP3 models in **a** January and **b** July for the area-weighted global mean (*asterisks*), Arctic mean (*diamonds*) and Antarctic mean (*triangles*). The Arctic and Antarctic means are area-weighted means poleward of 60°. The *vertical dashed lines* show the multi-model mean weight ($1/n$)

The precision of the estimated mean response is quantified based on its variance. For independent residual errors about the line of best fit, the variance of the mean response is

$$s_{\hat{y}}^2 = \mathrm{var}(\hat{y}_0) = \sigma_\varepsilon^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right] \quad (6)$$

where $\sigma_\varepsilon^2 = \mathrm{var}(\varepsilon)$ (see Draper and Smith 1998; p. 130). The second term in parentheses accounts for estimation error in the slope and grows quadratically with the distance of the observed present day value from the ensemble mean of the climate models. Equation (6) can be used to quantify the precision of the estimated response. For example, for normally distributed residual errors, the 95 % confidence interval in the mean response is given by $(\hat{y}_0 - 1.96 s_{\hat{y}}, \hat{y}_0 + 1.96 s_{\hat{y}})$ for sufficiently large ensemble size n. The 95 % prediction interval additionally takes into account the residuals about the mean response and is given by $\left(\hat{y}_0 - 1.96\sqrt{(s_{\hat{y}}^2 + \sigma_\varepsilon^2)}, \hat{y}_0 + 1.96\sqrt{(s_{\hat{y}}^2 + \sigma_\varepsilon^2)}\right)$. For the ensemble mean case with slope $\beta = 0$, the mean response is $\hat{y}_0 = \bar{y}$ with variance $s_{\hat{y}}^2 = \mathrm{var}(\hat{y}_0) = \sigma_y^2 n^{-1}$ (assuming independent y_i). For non-zero slope, the variance of the residuals about the regression line, σ^2, will be smaller than the variance of residuals from the ensemble mean response and this can then lead to smaller variance in the mean response from the regression approach (i.e. more precise climate change projection). The ensemble mean approach generally makes the simplest assumption that the model responses y_i are independent and identically distributed about the same mean. However, model responses are not unconditionally independent if the responses are state-dependent, or if the individual model errors are correlated with one another (Knutti et al. 2010; Weigel et al. 2010; Stephenson et al. in review). This violation is one of the main motivations for using ensemble regression where one makes the weaker, and more justifiable, assumption that residuals of the responses from the regression fit are independent. Allowing for positive correlations between model responses would inflate the variance of the ensemble mean projections. Hence, the comparison of results in this paper is conservative in that it compares the precision of projections derived from ensemble regression to overestimates of the precision of projections derived from the ensemble mean approach.

It should be noted that our regression equations are estimated using only the model data and so observational error will have no effect on the estimated slope and intercept. However, the projection of the future mean value does potentially depend on the choice of observation since it is the linear statistical prediction evaluated at the observed present day value. The sensitivity to the choice of dataset is considered in the next section.

Because of the small number of climate models, it is important to test how much influence each model is having on the mean response. We investigate this by calculating the leverage for each model (see "Appendix"), which helps identify overly influential models that can then be removed if desired. Figure 4 shows spatial averages [global, NH winter (Fig. 4a) and SH winter (Fig. 4b)] of grid point leverages of the CMIP3 models. One rule of thumb for labelling cases as "high leverage" is if the leverage exceeds $3p/n$ where p is the number of predictor variables and n is the sample size (Hoaglin and Kempthorne 1986). For our data example, $n = 24$ and $p = 1$ so the threshold for high leverage is $3/24 = 0.125$. In terms of global averages, none of the models have a particularly large leverage (i.e. there are no particularly influential outliers). However, over the Arctic in winter model 10 has a much larger leverage than the other models (Fig. 4a). This is the result of the unrealistically small poleward ocean heat transport at mid-latitudes in this model, which results in much larger sea ice extents than observed in both hemispheres (Arzel et al. 2006). Since it is an influential outlier at mid-to-high latitudes with a clear physical deficiency, model 10 is omitted from the main results for both the Arctic and Antarctic.

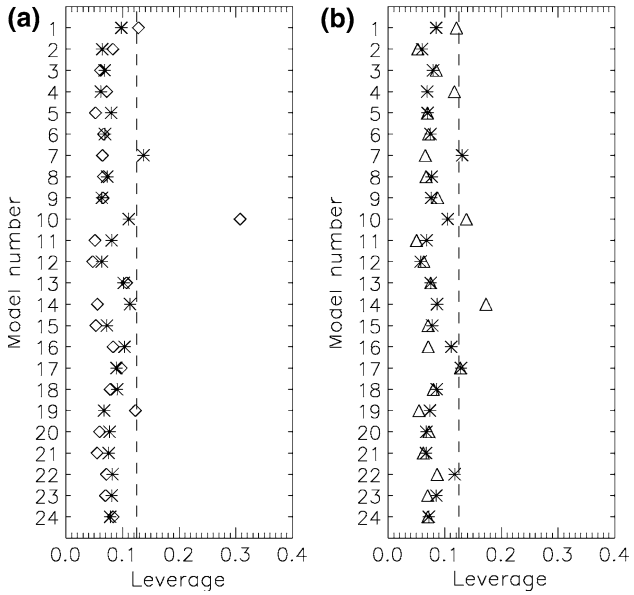

Fig. 4 Area-averaged leverages for each of the CMIP3 models in **a** January and **b** July for the area-weighted global mean (*asterisks*), Arctic mean (*diamonds*) and Antarctic mean (*triangles*). The Arctic and Antarctic means are area-weighted means poleward of 60°. The *vertical dashed lines* show the rule of thumb value of $3p/n$ for labelling cases as high leverage (Hoaglin and Kempthorne 1986)

The leverage-based approach exploits the whole ensemble to identify potential outlier models rather than to reject models on individual performance based on ad hoc metrics. We also use cross validation and bootstrap resampling to investigate the sensitivity of our results to the choice of models—ideally, the mean response is not unduly sensitive to which particular subset of models we decide to choose.

2.4 Near-surface temperature data

The multi-model ensemble climate model dataset used is the Coupled Model Inter-comparison Phase 3 (CMIP3) dataset, which was compiled as part of the Intergovernmental Panel on Climate Change (IPCC) Fourth Assessment Report (AR4). Model projections based on emissions following the Special Report on Emissions Scenarios (SRES) A1B scenario are used here. In terms of global temperature, the SRES A1B scenario is about the middle of the range of changes projected by the different SRES scenarios. Historical forcing runs (20C3M) were used for the present day time slice. Changes over the twenty-first century are defined as differences between a future time slice (2069–2098) and a present day time slice (1970–1999). A total of 24 different climate models were considered in this study and are listed in Table 1. For each model the average of all available ensemble members was taken and used to calculate mean surface air temperature (T_s; CMIP3 variable name 'tas') over both time slices. Before analysis the climate model datasets were bi-linearly interpolated onto the UKMO-HadCM3 horizontal grid (2.5° latitude × 3.75° longitude). A sensitivity test of results using the CSIRO-Mk3.5 grid (1.9° × 1.9°) showed negligible differences (see Sect. 3.2).

The European Centre for Medium-Range Weather Forecasts (ECMWF) ERA-40 re-analysis dataset was used for observations. Over the polar regions (particularly Antarctica) there is a dramatic increase in accuracy of re-analysis datasets after the introduction of widespread satellite temperature retrievals in late 1978 (Hines et al. 2000; Marshall and Harangozo 2000; Renwick 2004; Sterl 2004). Therefore the period used to calculate the present day mean climate from ERA-40 was 1979–1999. Estimates of the errors in ERA-40 fields are not provided by ECMWF. However, various comparisons between ERA-40 fields and in situ observations at high latitudes have been conducted. These show post-1979 T_s biases in ERA-40 of approximately ±1 °C over both the Arctic and Antarctic (Bromwich and Fogt 2004; Bromwich et al. 2007; Tjernstrom and Graversen 2009; Brodeau et al. 2010). We repeated ER projections using the recently-released ERA-Interim dataset (Dee et al. 2011) instead of ERA-40 and obtained very similar results (not shown). Internal climate variability will also introduce sampling error to the estimation of present day background climate. The standard error of January 20-year averages of T_s was estimated to be approximately 0.5–1.5 °C near the Arctic sea ice edge from a pre-industrial control run of ECHAM5/MPI-OM (not shown). The total range of uncertainty in the ERA-40 data is therefore an order of magnitude smaller than the inter-model range in T_s of approximately 20 °C over sea ice in winter (Fig. 2c, d).

3 Results

3.1 Arctic and Antarctic temperature projections

Figure 5 shows projections of twenty-first century grid point T_s change over the Arctic in winter (January), with the regressions for ER calculated at each grid point. Model 10 was omitted as it was found to be an influential outlier (see Sect. 2.3). The EM method (Fig. 5a) shows the largest warming over the Arctic Ocean. The warming is particularly large over Hudson Bay, the Chuckchi Sea and the Barents Sea. The projections derived from the ER method show a broadly similar pattern (Fig. 5b), but with key differences (Fig. 5c). The most notable difference is less warming over the Barents Sea by approximately 3 °C. Here there is a particularly strong gradient in warming, which is shifted further east in ER. There is also significantly less warming over parts of the northern boundary of the Pacific (significance indicated by hatching). In addition there is significantly more warming in a region south of Newfoundland in the northwest Atlantic. Figure 5d shows that these differences are associated with biases in the CMIP3 ensemble average present day climatology, which in particular shows a large cold bias of more than 10 °C over the Barents Sea. Most of the models have too much sea ice in the Barents Sea (Arzel et al. 2006) and therefore as this retreats in the future they give unrealistically large warming in a region that in reality has an ice-free present day climate. The ER method therefore adjusts for this unrealistically large warming over the Barents Sea, which is a clear improvement over the EM method. Over the eastern region of the Arctic Ocean the ER method gives more warming, but not significantly more. This is a consequence of the change in sign of the regression slope poleward of the ice edge region that was seen in Fig. 2. Since ER gives more warming in some places and less in others, for the whole Arctic region (all grid points ≥60° N) differences in warming are negligible with an area-weighted warming of 6.9 °C given by the EM approach and ER giving 7.0 °C.

Over southern high latitudes in winter (July) there is in general less warming than over the Arctic in winter (Fig. 6). As was found for the Arctic, there are regions with significant differences between the ER and EM methods. The ER method gives warming of up to 7 °C over a region to the northwest of the Weddell Sea at approximately 62° S, 5° W (Fig. 6b). This is approximately 2 °C more than estimates based on the EM method (Fig. 6c). There is a large region of significantly less warming extending westwards from the tip of the Antarctic Peninsula (centred on ~60° S, ~90° W). A smaller region of reduced warming is apparent at around 60° S, 110° E over the Southern Ocean. As was found in the Arctic winter these differences coincide with regions of large bias in the

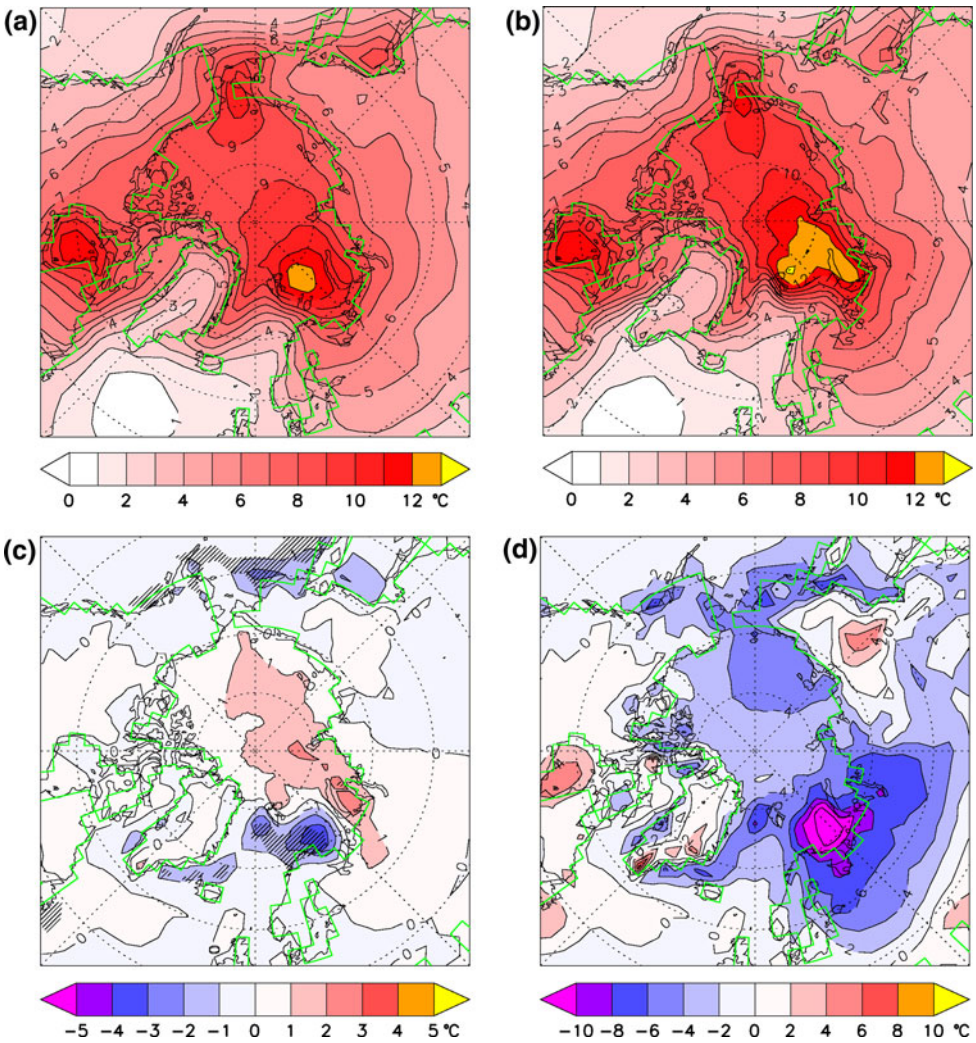

Fig. 5 Estimates of January near surface temperature change over the twenty-first century from **a** the EM method and from **b** the ER method. Model 10 is excluded and ERA-40 is used for the observed near-surface temperature. **c** The difference between **b** and **a**, with locations of significant difference indicated by *hatching*. A difference is considered significant if the EM-derived projection lies outside the 95 % confidence interval of the ER-derived projection (e.g. Fig. 2d). **d** The difference between the present day climatology in the CMIP3 ensemble mean and ERA-40

present day climatology of the CMIP3 ensemble average (Fig. 6d). Again this demonstrates the advantage of using the ER method over the EM method in taking account of unrealistic projected change related to present day bias.

In Fig. 7, the 95 % prediction interval is used to quantify the precision of projections derived from the ER and EM methods. Both methods show the largest prediction intervals over regions of sea ice, up to 10 °C over the Arctic. Compared to the EM method, the ER method gives prediction intervals that are approximately 30 % smaller over the Sea of Okhostk, the Bering Sea and the Labrador Sea (Fig. 7c). Over the Arctic Ocean reductions in prediction interval are smaller due to the weaker relationships there. The most dramatic reductions in prediction interval occur in the vicinity of the winter sea ice edge around Antarctica (Fig. 7f). In particular, reductions of 50 % extend across a sector of the Southern Ocean between 30° W and 90° E. At the northern part of the Antarctic Peninsula the ER method gives reductions in prediction interval of up to 30 %. This is a significant improvement on the weighted projections shown in Bracegirdle et al. (2008), which showed no reductions in inter-model spread compared to unweighted projections around Antarctica. At lower latitudes the ER method once again effectively reverts to the EM method with ratios of approximately one.

3.2 Cross validation errors

Are the improvements in precision seen in the ER method robust to other ways of quantifying precision? To answer this question, cross validation tests of precision were conducted. Cross validation has been used in previous studies (Raisanen et al. 2010; Abe et al. 2011) and therefore allows for more direct comparison with the results in this paper. In cross validation, 'observations' are taken from a validation model and the remaining models in the ensemble are used to make a projection that can be verified against that validation model. This is repeated with each model in turn

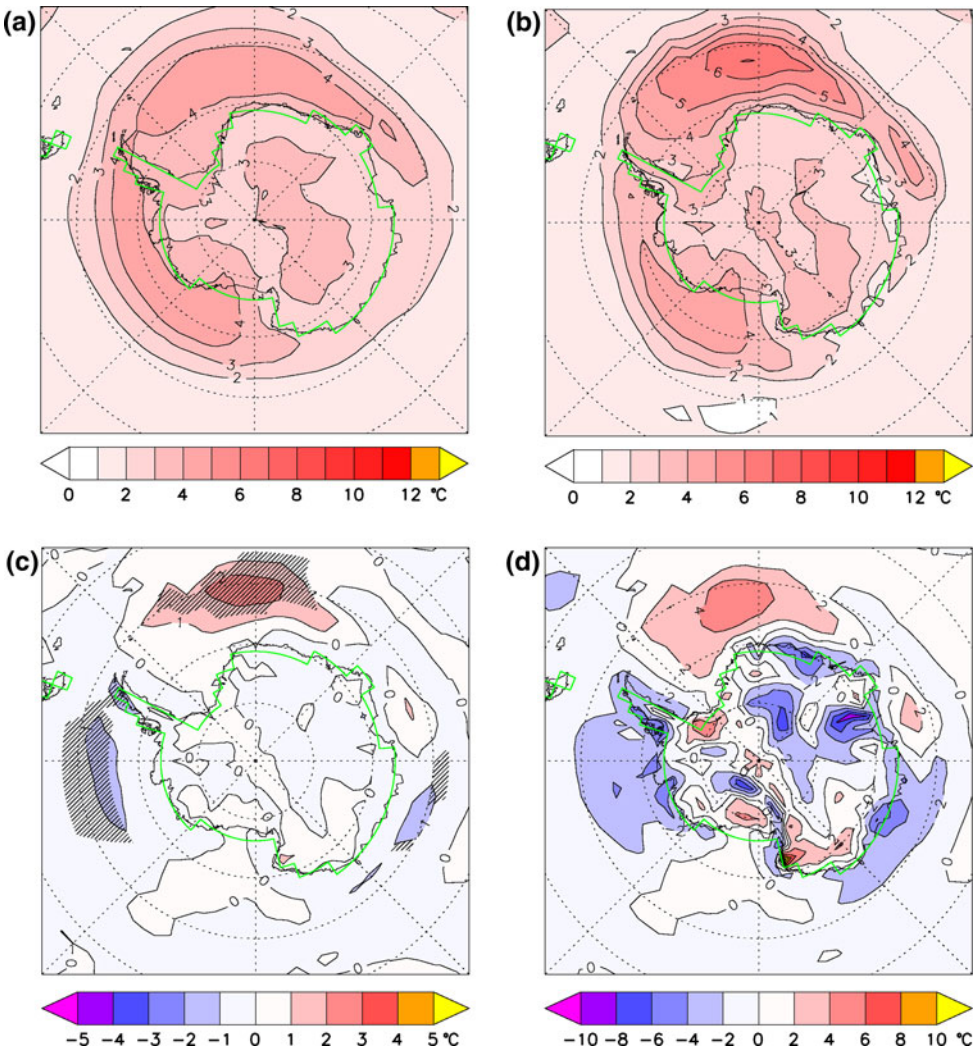

Fig. 6 As in Fig. 5 but for Antarctic winter (July)

taken as the validation model. The quadratic mean from cross validation squared errors across all the models was then calculated for projections derived from the ER and EM methods. The root mean squared error (RMSE) for January and July is shown in Fig. 8. The ratio of the ER to the EM RMSE gives almost identical results to the ratio of prediction intervals (compare Figs. 7, 8). The global plots in Fig. 8 demonstrate that in their respective summer seasons there is little difference in performance between the two methods in the Arctic and Antarctic.

Away from the winter sea ice edge there are some locations of RMSE reduction that may warrant further study (Fig. 8e, f). In the Americas, both southern Brazil (January only) and a region straddling southern Mexico (January and July) show RMSE reductions of $\sim$10 %. Over South Africa there are also reductions of more than 10 % in January. Over SE Asia there is a region of RMSE reductions of 10–20 % in July. Over the southern Antarctic Peninsula in summer (January) reductions of up to 30 % are evident. These changes are relatively small, but other predictor variables might give better results.

The reductions in cross validation error relative to the EM method are slightly larger than error reductions achieved by the weighting method of Raisanen et al. (2010). They found a global all-month reduction in quadratic mean squared error (MSE) of 4.7 % (from an EM MSE of 1.157 °C) when grid point T_s was used for predictor and predictand. Using the same 23 models (all models listed in Table 1 apart from model 16) and CMIP3 run numbers (only 'run1') as Raisanen et al. (2010) we found a reduction of 6.0 % (from an EM MSE of 1.187 °C). The slightly larger EM MSE found here could be a consequence of the different grids used for combining the model data. Use of the CSIRO-Mk3.5 grid reduced the MSE error by 0.9 % for both ER and EM, demonstrating a small sensitivity to the choice of grid, but with no impact on the relative performance of ER compared to EM.

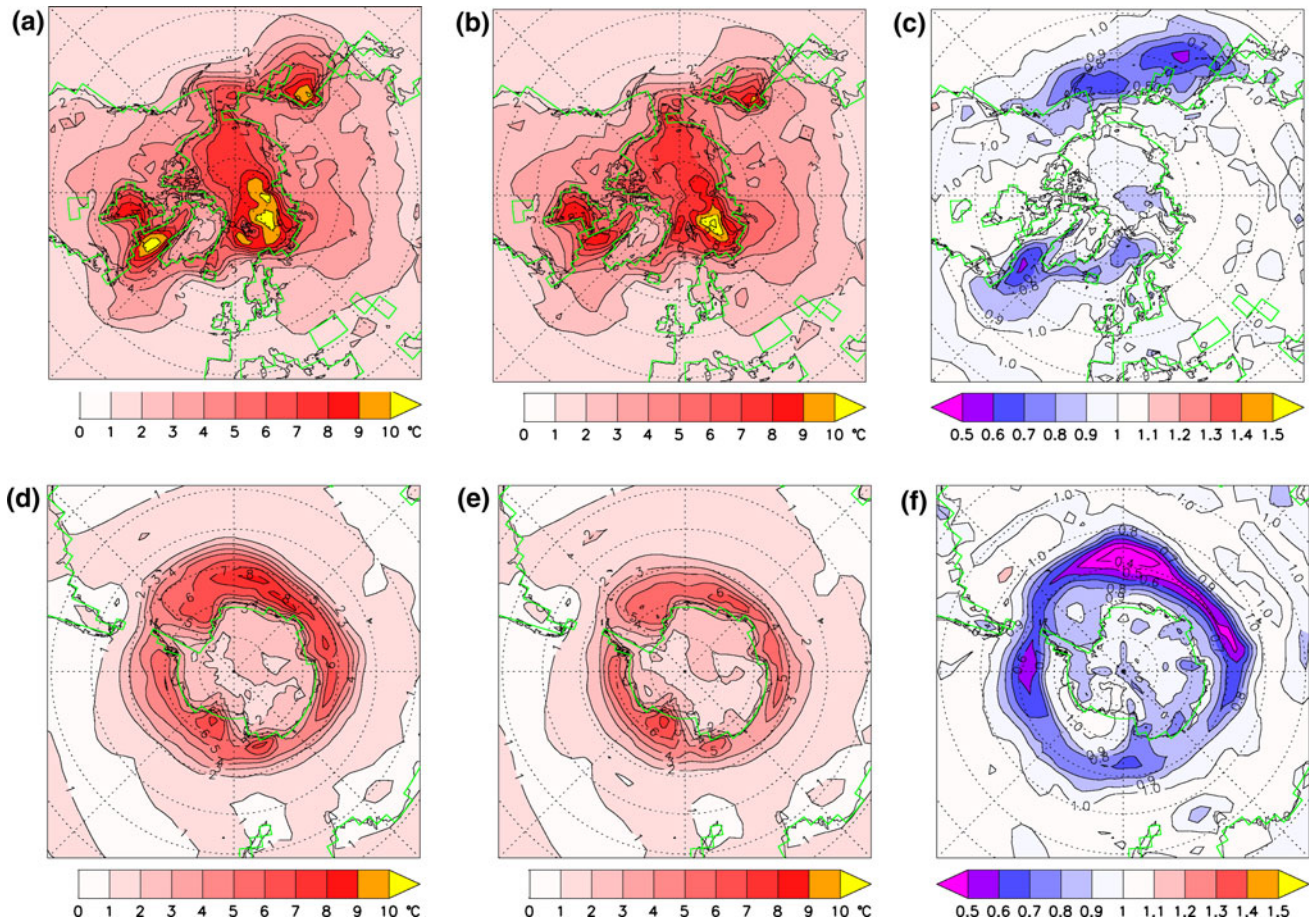

Fig. 7 Width of the 95 % prediction interval for **a**, **d** the EM method, **b**, **e** the ER method and **c** the ratio of **b** to **a**, and **f** the ratio of **e** to **d**. The *upper row* **a–c** shows results for January and the *lower row* **d–f** shows results for July

Compared to Abe et al. (2011), reductions in ER cross validation error in projected T_s change are similar over the Arctic Ocean (5–10 %) but larger along the Arctic ice edge. At mid-to-high southern latitudes there is a large contrast in results, with increases in RMSE of 10–15 % reported by Abe et al. (2011) compared to the large reductions found here. The reasons for this are not clear, but it is likely that their exclusion of latitudes south of 60° S in the domain used for calculation of SVD modes is a contributory factor.

Since the above MSE results are based on only 'run1' from each model, a further sensitivity test was conducted to assess whether including multiple ensemble members from each model has a significant impact on the results. If, as elsewhere in this paper, a mean of all ensemble members is used for each model, the reduction in ER MSE compared to EM MSE increases slightly to 6.4 %. This shows a slight benefit of using the mean of multiple ensemble members for each model, which will remove some of the sampling error in extracting the background climate using 30-year means from a single ensemble member.

3.3 Sensitivity to ensemble size

Would the addition of more models result in further improvements in ER precision? How sensitive are the ER results to the choice of models? In this section these important questions are answered using a bootstrap resampling method in which cross validation is applied to randomly resampled sub-ensembles of different sizes. This was performed as follows: (1) A validation model was selected at random from 23 models (model 10 omitted). (2) From the remaining 22 models, sub-ensembles of m models ($m = 2, 3, 4, \ldots, 22$) were then selected at random (with replacement allowed). For each m, (1) and (2) were repeated 1,000 times.

This resampling approach was used for area-weighted averages of grid point projections over the Barents Sea in January (region B) and a region to the northeast of the Weddell Sea in July (region W) (Fig. 9). These are regions of larger ER/EM difference and larger model bias in Figs. 5, 6. For convenience, the RMSE values shown in Fig. 9 are normalised by $\mathrm{RMSE}_{\mathrm{EM}}$ for m of 22. For region

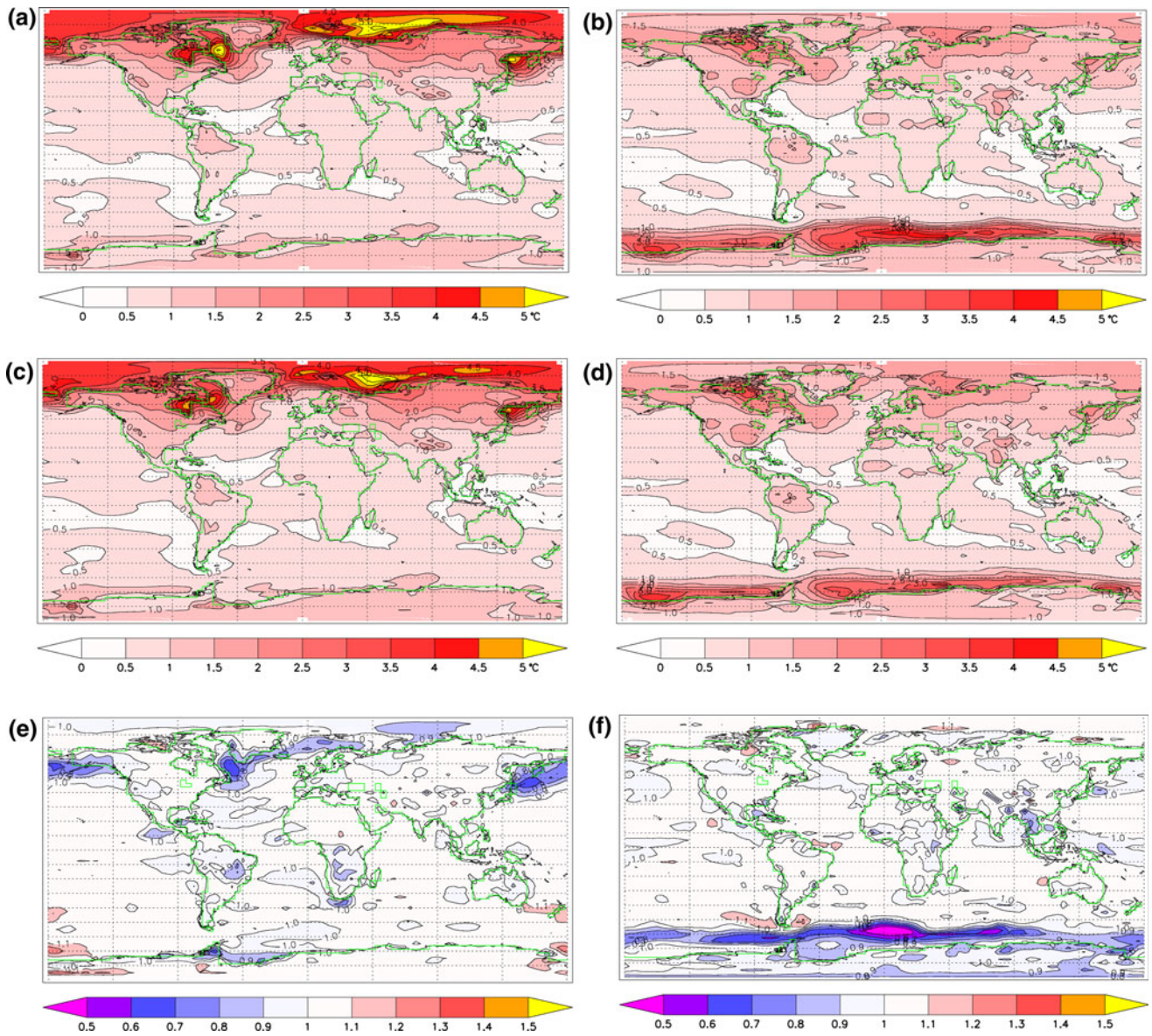

Fig. 8 Cross validation errors averaged over 23 CMIP3 models (model 10 excluded). The *upper row* **a**, **b** shows the RMSE for the EM method, the *middle row* **c**, **d** shows the RMSE for the ER method and the *bottom row* **e**, **f** shows data from the *middle row* divided by data from the *upper row*. The *left column* (**a**, **c**, **e**) shows results for January and the *right column* (**b**, **d**, **f**) shows results for July

B in January the median $RMSE_{ER}$ is very large for small m due to singularities in the regression when the same model is chosen for all members of the sub ensemble. However, this quickly reduces with increasing m. For m larger than 6 the median $RMSE_{ER}$ (solid line) is smaller than the median $RMSE_{EM}$ (dashed line). The decrease in median $RMSE_{ER}$ with increasing m ceases for m of approximately 12. The upper and lower quartiles also show decreases, although smaller for the upper quartile. In region W an m of only 3 is sufficient to give smaller median $RMSE_{ER}$ values compared to $RMSE_{EM}$ (Fig. 9b). This is due to the very strong relationship over region W. An m of only 6 is large enough to achieve close to maximum benefit from the ER method, although the lower quartile decreases slowly up to an m of approximately 14. By coincidence the upper quartile of the $RMSE_{ER}$ over region W is almost identical to the median $RMSE_{EM}$. As in the previous section, it was found that the use of only one ensemble member from each model had a negligible effect on the results.

3.4 Quadratic regression

Can the ER precision be increased further by extending the linear regression model to a higher-order (quadratic) model? To answer this question a cross validation test for

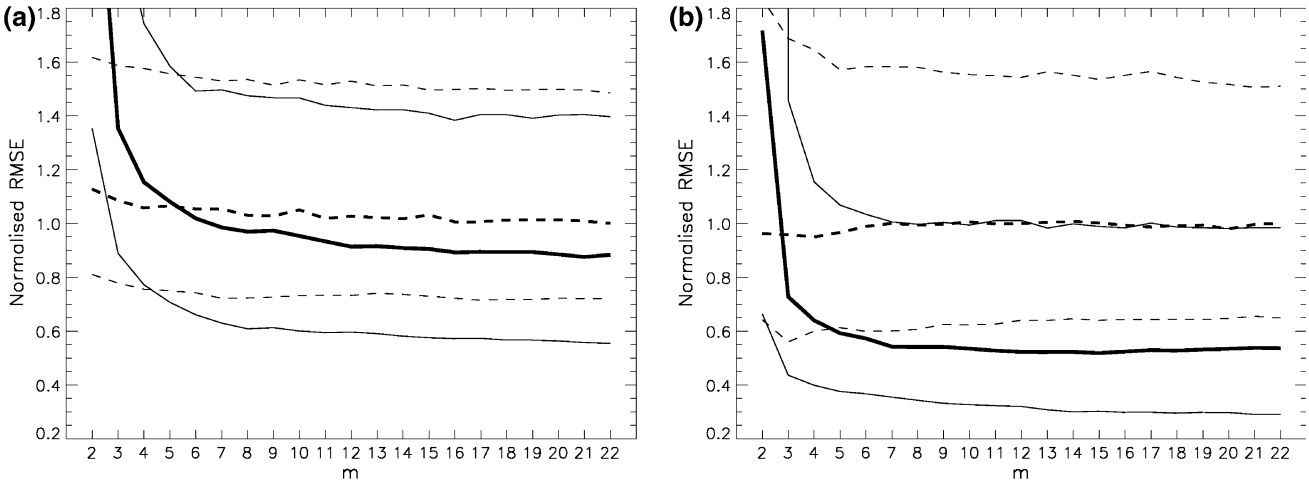

Fig. 9 Dependence of area-weighted cross validation RMSE error on number of randomly selected models (with replacement) (m) for **a** region B in January and **b** region W in July. For each m, 1,000 randomly selected model combinations were tested. *Bold lines* show the median RMSE from these combinations with the *upper* and *lower* quartiles shown by the *thin lines*. Results from the ER method are shown by the *solid lines* and *dashed lines* show results from the EM method. The RMSE errors are normalised in each plot by the median EM RMSE with $m = 22$. Model 10 was omitted. Region B is defined as an area-weighted spatial average over 70° N–80° N and 10° E–50° E and region W is defined as 55° S–65° S and 20° W–20° E

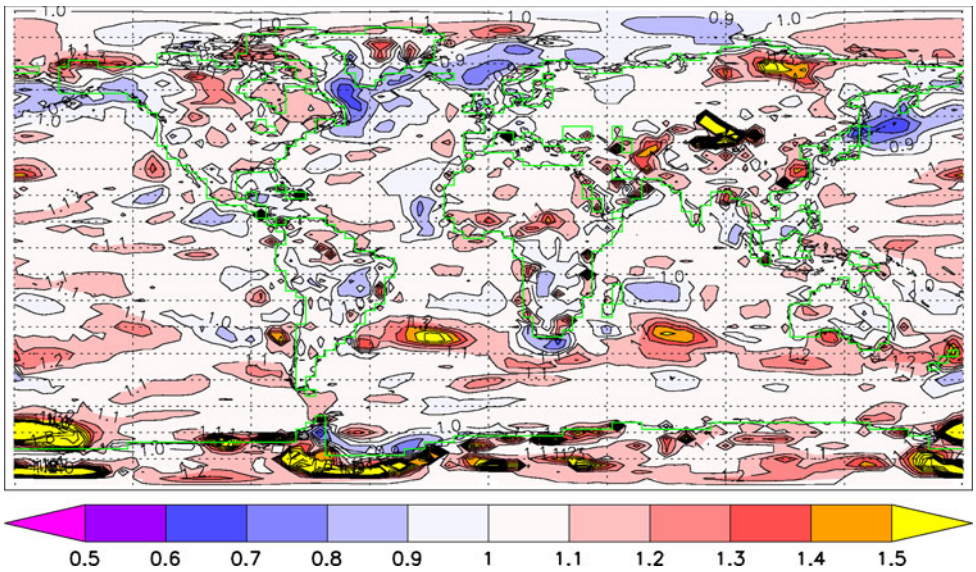

Fig. 10 As in Fig. 8e, but with a quadratic regression model

January was conducted with a quadratic regression model (Fig. 10). This shows almost identical RMSE reductions in the locations at which reductions occur in the linear case (compare with Fig. 8e). Similar results were also found for July (not shown). Linear regression therefore seems like a good description of the relationships in T_s near the winter sea ice edge. Figure 10 also indicates that the linear assumption is a more suitable choice over other parts of the globe. There are some regions of very large cross validation RMSE that are associated with the higher sensitivity of the quadratic model to outliers. The example in Fig. 11 illustrates this by comparing the effect of excluding model 1 on linear and quadratic fits at 65° N, 105° W. In the case of model 1 being chosen as the validation model, the quadratic fit in Fig. 11b would be used to estimate the change projected by model 1 (Fig. 11a). Clearly this would give a much larger warming than actually occurs in model 1. The linear fit is much less sensitive to the exclusion of model 1 (Fig. 11c, d).

4 Conclusions

This study has presented and tested an ensemble regression methodology, which gives near-surface temperature projections over the polar regions that are more precise than

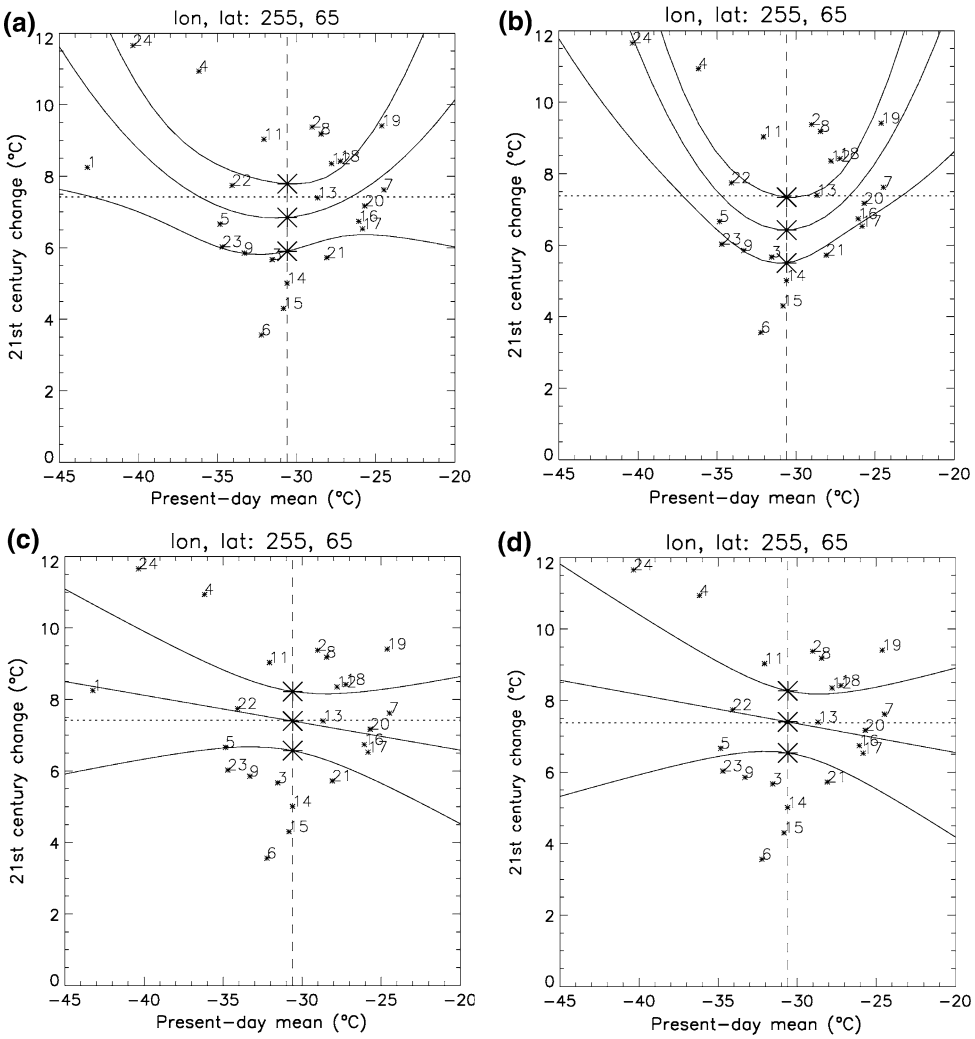

Fig. 11 Scatter plots in a similar format to those shown in Fig. 2, but for 65° N, 105° W in January. The *top row* **a**, **b** shows results from a quadratic regression and the *bottom row* **c**, **d** from a linear regression. Model 1 is included in the *left column* (**a**, **c**) and omitted from the *right column* (**b**, **d**)

those derived using equal-weight ensemble means. In addition to improving on the performance of previous weighting methods (Bracegirdle et al. 2008; Raisanen et al. 2010; Abe et al. 2011), the ER method avoids the additional complexity, computational expense and statistical uncertainty associated with the calibration of explicit model weights. The method has been successfully applied to CMIP3 gridded multi-model ensemble data to produce twenty-first century wintertime surface temperature projections under the SRES A1B emissions scenario.

Over the Arctic in January, the ER method gives less warming than the EM approach along the sea ice edge due to a widespread negative bias in present day surface temperature in the CMIP3 models. Most notably the results show 3 °C less warming over Barents Sea (∼7 °C compared to ∼10 °C) and 2 °C less warming over the Bering Sea (∼5 °C compared to ∼7 °C). For the whole Arctic region (all grid points ≥60° N) the differences in warming are negligible, with an area-weighted warming of 6.9 °C given by the EM approach and ER giving 7.0 °C. In addition, the ER method gives more precise projections near the sea ice edge, with reductions in projection uncertainty of approximately 30 % over the Sea of Okhostk, Bering Sea and Labrador Sea.

For the Antarctic in July, the ER method gives 2 °C more warming than the EM method (∼7 °C compared to ∼5 °C) over Southern Ocean across the Greenwich Meridian. In contrast, there is 1 °C less warming along a sector of the Southern Ocean that extends from the northern Antarctic Peninsula to approximately 120° W. Probably more important are the dramatic increases in precision around the SH winter sea ice edge. Projection uncertainty with the ER approach is almost half that of the EM approach over the Southern Ocean between 30° W to 90° E and up to 30 % over the northern Peninsula. An implication of these results is that the current maximum of winter warming over the Antarctic Peninsula and into West Antarctica (Chapman and Walsh 2007; Thomas et al. 2009; Steig et al. 2009) is not likely to continue under the SRESA1B scenario. Precise projections of the climate of

the Antarctic Peninsula and Greenland are important since these are regions in which dramatic changes in ice mass balance have recently been observed in connection with atmospheric change (Marshall et al. 2006; Pritchard et al. 2009). There are also implications for changes in extreme weather, in particular intense mesoscale winter cyclones (polar lows). Less surface warming over the Barents Sea would act to reduce the projected local increase in severity of polar lows in the future (Kolstad and Bracegirdle 2008; Zahn and von Storch 2010).

The ER methodology encompasses three important elements: (1) linear regression of climate change response on present day climate, (2) an assessment of leverage to identify influential outliers in the regression and (3) the use of cross validation to determine the point at which errors stop decreasing with increasing ensemble size (or whether a larger ensemble is required). Where there is no strong relationship between the response and present day climate, the ER approach yields the same projections as one would find using EM. Our approach has the advantage that it is simpler and less subjective than that of Raisanen et al. (2010) and is more robust to overly influential models and individual model biases than the approach used in Boe et al. (2009). The linear model assumption was found to be a good description of inter-model relationships near the winter sea ice edge. Cross validation errors using a quadratic regression model were similar to those calculated using linear regression. Additionally quadratic regression produced large errors at some locations due to a stronger sensitivity to outliers. The second element, leverage, is a powerful tool which enables the identification of outliers that have a strong influence on the regression. In this study, the leverage showed a clear influential outlier at high latitudes in winter (Fig. 2). This model (model 10) singularly accounts for a large part of previously documented inter-model relationships between present day and future T_s over NH mid-latitude oceans (e.g. Raisanen 2007; Knutti et al. 2010). In this case the leverage test was effective in identifying a model that has previously documented problems with sea ice extent (e.g. Arzel et al. 2006; Connolley and Bracegirdle 2007) and suggests that it is a useful tool for quickly flagging potentially problematic models in new multi-model datasets.

Another challenge when assessing new multi-model datasets is the generally small number of models. In particular a small ensemble size causes problems in calibrating model weights (Knutti et al. 2010; Raisanen et al. 2010). However, here it was found that where relationships are strong near the SH sea ice edge an ensemble size of only approximately 6 is sufficient to give cross validation error statistics that are nearly invariant under further increases in ensemble size. Over the Barents Sea, where the relationship is weaker, a larger ensemble size of approximately 12 is required. The required ensemble size therefore varies depending on the location considered. The results may also be specific to the CMIP3 dataset and variable considered.

For simplicity in introducing the ER method we have focussed on time-mean grid point T_s for both predictor and predictand. In principle any variable or combination of variables could be used as the predictor. Raisanen et al. (2010) assessed a range of predictors for future change in T_s and found for instance that present day T_s variability gave better results globally than the climatological mean. It may therefore be possible to achieve further increases in precision with different choices of predictor variable in ER. In considering other predictors or predictands a clear physical understanding of how the variables are related is important. The future/present day correlations that have been found in previous work are generally related to the spatial movement of physical features under future emissions scenarios, most notably the sea ice edges (Raisanen 2007; Knutti et al. 2010) and mid-latitude storm tracks (Whetton et al. 2007; Giorgi and Coppola 2010; Kidston and Gerber 2010). Position error of these features in simulations of present day climate will clearly have an impact on predicted future change of related parameters as they change position. It may therefore be useful to apply the ER method to predictand variables relating to storm tracks, particularly over the Southern Hemisphere. It is intended that a more extensive analysis of other variables and seasons will be conducted using the CMIP5 database once sufficient data are available.

Acknowledgments This study is part of the British Antarctic Survey Polar Science for Planet Earth Programme. It was funded by the Natural Environment Research Council. Two anonymous authors are thanked for their useful comments, which helped to significantly improve the manuscript. We acknowledge the modeling groups for making their simulations available for analysis, the Program for Climate Model Diagnosis and Inter-comparison (PCMDI) for collecting and archiving the CMIP3 model output, and the WCRP's Working Group on Coupled Modelling (WGCM) for organizing the model data analysis activity. The WCRP CMIP3 multimodel data set is supported by the Office of Science, U.S. Department of Energy. The European Centre for Medium Range Weather Forecasting are thanked for providing the ERA-40 and ERA-Interim datasets.

Appendix: Influential items in regression

The regression model in (2) can be written in matrix form as $\mathbf{Y} = \mathbf{X}b + \varepsilon$. $\mathbf{Y}$ is a column vector of n climate model simulations of the climate response, $\mathbf{X}$ is the design matrix with n rows and 2 columns: the first column is filled with ones and the second column contains the present day values $(x_1, \ldots, x_n)'$ from each climate model. For quadratic regression, a third column with $(x_1^2, \ldots, x_n^2)'$ is added to $\mathbf{X}$. $\boldsymbol{\varepsilon}$ is a column vector of n residuals.

The regression model parameters are given by the two row column vector $b = (\mu, \beta)'$, which is estimated to be $\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ using ordinary least squares estimation (Draper and Smith (1998), p 125). The estimated values of climate model responses are then given by $\hat{\mathbf{Y}} = \mathbf{X}\hat{b} = \mathbf{X}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}) = \mathbf{HY}$, where $\mathbf{H} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is known as the *'hat matrix'* (Draper and Smith (1998), p205). The *i*th diagonal element of the hat matrix H_{ii} is called the *leverage* of the *i*th item, and helps to quantify how influential each item is on the overall fit (in our case, the items are the climate model simulations). Items having large leverage are known as *influential* items. One rule of thumb for labelling cases as "high leverage" is if the leverage exceeds $3p/n$ where p is the number of predictor variables and n is the sample size (Hoaglin and Kempthorne 1986). Influential items can unduly modify regression estimates especially if they are also outlier points with large residuals.

References

Abe M, Shiogama H, Nozawa T, Emori S (2011) Estimation of future surface temperature changes constrained using the future-present correlated modes in inter-model variability of CMIP3 multi-model simulations. J Geophys Res Atmos 116:D18104. doi:10.1029/2010jd015111

Arzel O, Fichefet T, Goosse H (2006) Sea ice evolution over the 20th and 21st centuries as simulated by current AOGCMs. Ocean Model 12:401–415. doi:10.1016/j.ocemod.2005.08.002

Boe JL, Hall A, Qu X (2009) September sea-ice cover in the Arctic Ocean projected to vanish by 2100. Nat Geosci 2(5):341–343. doi:10.1038/ngeo467

Bracegirdle TJ, Connolley WM, Turner J (2008) Antarctic climate change over the twenty first century. J Geophys Res Atmos 113(D3). doi:10.1029/2007jd008933

Brodeau L, Barnier B, Treguier AM, Penduff T, Gulev S (2010) An ERA40-based atmospheric forcing for global ocean circulation models. Ocean Model 31(3–4):88–104. doi:10.1016/j.ocemod.2009.10.005

Bromwich DH, Fogt RL (2004) Strong trends in the skill of the ERA-40 and NCEP-NCAR reanalyses in the high and midlatitudes of the Southern Hemisphere, 1958–2001. J Clim 17(23):4603–4619. doi:10.1175/3241.1

Bromwich DH, Fogt RL, Hodges KI, Walsh JE (2007) A tropospheric assessment of the ERA-40, NCEP, and JRA-25 global reanalyses in the polar regions. J Geophys Res Atmos 112(D10):D10111. doi:10.1029/2006jd007859

Chapman WL, Walsh JE (2007) A synthesis of Antarctic temperatures. J Clim 20(16):4096–4117. doi:10.1175/jcli4236.1

Christensen JH, Hewitson B, Busuioc A, Chen A, Gao X, Held I, Jones R, Kolli RK, Kwon W-T, Laprise R, Rueda VM, Mearns L, Menéndez CG, Räisänen J, Rinke A, Sarr A, Whetton P (2007) Regional climate projections. In: Solomon S, Qin D, Manning M, Chen Z, Marquis M, Averyt KB, Tignor M, Miller HL (eds) Climate change 2007: the physical science basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. C. U. Press, Cambridge

Connolley WM, Bracegirdle TJ (2007) An antarctic assessment of IPCC AR4 coupled models. Geophys Res Lett 34(22). doi:10.1029/2007gl031648

Dee DP, Uppala SM, Simmons AJ, Berrisford P, Poli P, Kobayashi S, Andrae U, Balmaseda MA, Balsamo G, Bauer P, Bechtold P, Beljaars ACM, van de Berg L, Bidlot J, Bormann N, Delsol C, Dragani R, Fuentes M, Geer AJ, Haimberger L, Healy SB, Hersbach H, Holm EV, Isaksen L, Kallberg P, Kohler M, Matricardi M, McNally AP, Monge-Sanz BM, Morcrette JJ, Park BK, Peubey C, de Rosnay P, Tavolato C, Thepaut JN, Vitart F (2011) The ERA-Interim reanalysis: configuration and performance of the data assimilation system. Q J R Meteorol Soc 137(656):553–597. doi:10.1002/qj.828

Draper NR, Smith H (1998) Applied regression analysis, 3rd edn. Wiley, New York

Giorgi F, Coppola E (2010) Does the model regional bias affect the projected regional climate change? An analysis of global model projections. Clim Chang 100(3–4):787–795. doi:10.1007/s10584-010-9864-z

Giorgi F, Mearns LO (2002) Calculation of average, uncertainty range, and reliability of regional climate changes from AOGCM simulations via the "reliability ensemble averaging" (REA) method. J Clim 15:1141–1158

Greene AM, Goddard L, Lall U (2006) Probabilistic multimodel regional temperature change projections. J Clim 19(17):4326–4343

Hall A, Qu X (2006) Using the current seasonal cycle to constrain snow albedo feedback in future climate change. Geophys Res Lett 33:L03502. doi:10.1029/2005GL025127

Hines KM, Bromwich DH, Marshall GJ (2000) Artificial surface pressure trends in the NCEP/NCAR reanalysis over the Southern Ocean and Antarctica. J Clim 12:3940–3952. doi:10.1175/1520-0442(2000)013<3940:ASPTIT>2.0.CO;2

Ho CK, Stephenson DB, Collins M, Ferro CAT, Brown SJ (2012) Calibration strategies: a source of additional uncertainty in climate change projections. Bull Am Meteorol Soc 93(1):21–26. doi:10.1175/2011bams3110.1

Hoaglin DC, Kempthorne PJ (1986) Influential observations, high leverage points, and outliers in linear regression: comment. Stat Sci 1(3):408–412

Holland MM, Bitz CM (2003) Polar amplification of climate change in coupled models. Clim Dyn 21:221–232

Kidston J, Gerber EP (2010) Intermodel variability of the poleward shift of the austral jet stream in the CMIP3 integrations linked to biases in 20th century climatology. Geophys Res Lett 37. doi:10.1029/2010gl042873

Knutti R, Furrer R, Tebaldi C, Cermak J, Meehl GA (2010) Challenges in combining projections from multiple climate models. J Clim 23(10):2739–2758. doi:10.1175/2009jcli3361.1

Kolstad EW, Bracegirdle TJ (2008) Marine cold-air outbreaks in the future: an assessment of IPCC AR4 model results for the Northern Hemisphere. Clim Dyn 30(7–8):871–885. doi:10.1007/s00382-007-0331-0

Mahlstein I, Knutti R (2011) Ocean heat transport as a cause for model uncertainty in projected arctic warming. J Clim 24(5):1451–1460. doi:10.1175/2010jcli3713.1

Marshall GJ, Harangozo SA (2000) An appraisal of NCEP/NCAR reanalysis MSLP viability for climate studies in the South Pacific. Geophys Res Lett 27:3057–3060

Marshall GJ, Orr A, van Lipzig NPM, King JC (2006) The impact of a changing Southern Hemisphere Annular Mode on Antarctic Peninsula summer temperatures. J Clim 19(20):5388–5404

Murphy JM, Sexton DMH, Barnett DN, Jones GS, Webb MJ, Collins M, Stainforth DA (2004) Quantification of modelling uncertainties in a large ensemble of climate change simulations. Nat 430:768–772. doi:10.1038/nature02771

Overland JE, Wang MY (2007) Future regional Arctic sea ice declines. Geophys Res Lett 34. doi:10.1029/2007GL030808

Pritchard HD, Arthern RJ, Vaughan DG, Edwards LA (2009) Extensive dynamic thinning on the margins of the Greenland and Antarctic ice sheets. Nat 461(7266):971–975. doi:10.1038/nature08471

Raisanen J (2007) How reliable are climate models? Tellus 59A:2–29. doi:10.1111/j.1600-0870.2006.00211.x

Raisanen J, Ruokolainen L, Ylhaisi J (2010) Weighting of model results for improving best estimates of climate change. Clim Dyn 35(2–3):407–422. doi:10.1007/s00382-009-0659-8

Renwick JA (2004) Trends in the Southern Hemisphere polar vortex in NCEP and ECMWF reanalyses. Geophys Res Lett 31:L027209. doi:10.1029/2003GL019302

Shindell D, Faluvegi G (2009) Climate response to regional radiative forcing during the twentieth century. Nat Geosci 2(4):294–300. doi:10.1038/ngeo473

Steig EJ, Schneider DP, Rutherford SD, Mann ME, Comiso JC, Shindell DT (2009) Warming of the Antarctic ice-sheet surface since the 1957 International Geophysical Year. Nat 457(7228):459–462

Sterl A (2004) On the (in)homogeneity of reanalysis products. J Clim 17(19):3866–3873

Stroeve J, Holland MM, Meier W, Scambos T, Serreze M (2007) Arctic sea ice decline: faster than forecast. Geophys Res Lett 34(9). doi:10.1029/2007gl029703

Tebaldi C, Rl Smith, Nychka D, Mearns LO (2005) Quantifying uncertainty in projections of regional climate change: a bayesian approach to the analysis of multimodel ensembles. J Clim 18:1524–1540. doi:10.1175/JCLI3363.1

Thomas ER, Dennis PF, Bracegirdle TJ, Franzke C (2009) Ice core evidence for significant 100-year regional warming on the Antarctic Peninsula. Geophys Res Lett 36:L20704. doi:10.1029/2009gl040104

Tjernstrom M, Graversen RG (2009) The vertical structure of the lower Arctic troposphere analysed from observations and the ERA-40 reanalysis. Q J R Meteorol Soc 135(639):431–443. doi:10.1002/qj.380

Walsh JE, Chapman WL, Romanovsky V, Christensen JH, Stendel M (2008) Global climate model performance over Alaska and Greenland. J Clim 21(23):6156–6174. doi:10.1175/2008jcli2163.1

Wang M, Overland JE, Kattsov V, Walsh JE, Zhang X, Pavlova T (2007) Intrinsic versus forced variation in coupled climate model simulations over the arctic during the twentieth century. J Clim 20(6):1093–1107

Watterson IG, Whetton PH (2011) Distributions of decadal means of temperature and precipitation change under global warming. J Geophys Res Atmos 116:D07101. doi:10.1029/2010jd014502

Weigel AP, Knutti R, Liniger MA, Appenzeller C (2010) Risks of model weighting in multimodel climate projections. J Clim 23(15):4175–4191. doi:10.1175/2010jcli3594.1

Whetton P, Macadam I, Bathols J, O'Grady J (2007) Assessment of the use of current climate patterns to evaluate regional enhanced greenhouse response patterns of climate models. Geophys Res Lett 34(14). doi:10.1029/2007gl030025

Zahn M, von Storch H (2010) Decreased frequency of North Atlantic polar lows associated with future climate warming. Nat 467(7313):309–312. doi:10.1038/nature09388

Zhang XD (2010) Sensitivity of arctic summer sea ice coverage to global warming forcing: towards reducing uncertainty in arctic climate change projections. Tellus Ser A Dyn Meteorol Oceanogr 62A(3):220–227. doi:10.1111/j.1600-0870.2010.00441.x