Performance of Pattern-Scaled Climate Projections under High-End Warming. Part I: Surface Air Temperature over Land

TIMOTHY J. OSBORN AND CRAIG J. WALLACE

Climatic Research Unit, School of Environmental Sciences, University of East Anglia, Norwich, United Kingdom

JASON A. LOWE[a] AND DAN BERNIE

Met Office Hadley Centre, Exeter, United Kingdom

(Manuscript received 14 November 2017, in final form 17 April 2018)

ABSTRACT

Pattern scaling is widely used to create climate change projections to investigate future impacts. We consider the performance of pattern scaling for emulating the HadGEM2-ES general circulation model (GCM) paying particular attention to ''high end'' warming scenarios and to different choices of GCM simulations used to diagnose the climate change patterns. We demonstrate that evaluating pattern-scaling projections by comparing them with GCM simulations containing unforced variability gives a significantly less favorable view of the actual performance of pattern scaling. Using a four-member initial-condition ensemble of HadGEM2-ES simulations, we infer that the root-mean-square errors of pattern-scaled monthly temperature changes over land are less than 0.25°C for global warming up to approximately 3.5°C. Some regional errors are larger than this and, for this GCM, there is a tendency for pattern scaling to underestimate warming over land. For warming above 3.5°C, the pattern-scaled projection errors grow but remain small relative to the climate change signal. We investigate whether patterns diagnosed by pooling GCM experiments from several scenarios are suitable for emulating the GCM under a high-end warming scenario. For global warming up to 3.5°C, pattern scaling using this pooled pattern closely emulates GCM simulations. For warming beyond 3.5°C, pattern-scaling performance is notably improved by using patterns diagnosed only from the high-forcing representative concentration pathway 8.5 (RCP8.5) scenario. Assessments of climate change impacts under high-end warming using pattern-scaling projections could be improved by using change patterns diagnosed from pooled scenarios for projections up to 3.5°C above preindustrial levels and patterns diagnosed from only strong forcing simulations for projecting beyond that. Similar findings are obtained for five other GCMs.

1. Introduction

Pattern scaling (PS) enables the generation of gridded, time-varying, climate change projections by combining the spatial climate-change responses of multiple general circulation models (GCMs) or Earth system models (ESMs) with a driving time series of global-mean temperature change ΔT_t. The GCM spatial climate responses (hereinafter patterns) can be diagnosed from any externally forced climate change simulations, such as the representative concentration pathway (RCP; van Vuuren et al. 2011) experiments from phase 5 of the Coupled Model Intercomparison Project (CMIP5; Taylor et al. 2012), provided the simulated response to the external forcing is large enough compared to GCM unforced variability. Multiple alternative ΔT_t values can then be prescribed, either as fixed specific warming levels (SWLs) or transient changes to explore a wide range of future scenarios and model uncertainties. Effectively, PS is an approximate physically based emulator for the more complex GCM behavior in terms of its geographical, seasonal, and multivariate response to

Denotes content that is immediately available upon publication as open access.

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/JCLI-D-17-0780.s1.

[a] Additional affiliation: Priestley Centre, University of Leeds, Leeds, United Kingdom.

Corresponding author: Timothy J. Osborn, t.osborn@uea.ac.uk

DOI: 10.1175/JCLI-D-17-0780.1

global anthropogenic forcing (Osborn et al. 2016). The appeal of PS is that climate projections for a wide pool of climate change scenarios, including combinations of GCMs and forcing scenarios not included in the training data, can be quickly generated to represent both GCM and scenario uncertainties. The approach is popular within integrated impact studies that couple socioeconomic and physical environmental prediction frameworks (e.g., Arnell et al. 2013; Warren et al. 2008; van Vuuren et al. 2006) or climate impact studies where running a suite of GCMs is unfeasible (e.g., Gosling and Arnell 2016; Ostberg et al. 2013; Warren et al. 2013).

Given its wide application, the ability of PS to emulate transient GCM simulations deserves further evaluation. The principal limitation of PS is that it encapsulates climate change responses only as a linear function of ΔT_t while the actual response of a GCM has features that evolve differently (e.g., Good et al. 2015). This includes both nonlinear changes and changes where there is a local change that either lags or leads the global average warming. Early validation of PS (Mitchell et al. 1999; Mitchell 2003) revealed that errors attributed to nonlinearities existed but were small compared to the size of the uncertainties arising from other factors (e.g., forcing scenarios, GCM choice, and climate variability). Recent analyses have found PS to be sufficient to approximate the greenhouse gas responses of the latest generation of GCMs (Heinke et al. 2013; Tebaldi and Arblaster 2014; Osborn et al. 2016) but have also found that limitations occur where strong regional differences in forcing exist (e.g., sulfate aerosols; Ishizaki et al. 2012), for climate variables with upper and/or lower bounds (e.g., cloud amount or precipitation) and for scenarios where the forcing stabilizes (Tebaldi and Arblaster 2014). Furthermore, in its simplest form, PS only represents changes in mean climate and not changes in internal climate variability. The PS tool considered here, ClimGen (Osborn et al. 2016), addresses some of these limitations—for instance, by applying nonlinear functions to precipitation and cloud cover and by superimposing observed anomalies onto the local scenarios and, in the case of precipitation, also transforming the anomalies to represent projected changes in interannual variability.

In this paper, we extend previous PS evaluations by focusing on three specific issues. First, we explore how the accuracy of PS (in terms of reproducing the transient GCM behavior) depends on the GCM simulation ensemble size—and thus on the accuracy with which we can diagnose the target climate change signal. Second, we examine the contribution of nonlinear climate system responses known to be present within the training GCM data. Third, and of importance to impact studies driven using PS data, is the performance of PS when approximating GCM behavior under high-end warming scenarios (i.e., SWLs of up to 6°C or those associated with the RCP8.5 concentration trajectory) when the patterns themselves have been diagnosed by pooling GCM simulations across multiple (usually weaker forcing) scenarios. We explore these issues using an ensemble of RCP simulations performed with the HadGEM2-ES GCM (Caesar et al. 2013) and confirm that the two key findings apply to five other CMIP5 GCMs as well (see section SM2 in the supplemental material).

2. Data and methods

a. The ClimGen pattern scaler

In its simplest form, PS estimates the future change in a climate variable V at a spatial grid cell at some time t in the future by

$$\Delta V_t = \alpha \Delta T_t, \tag{1}$$

where ΔT_t is the change in annual global-mean temperature relative to a preidentified baseline. Coefficient α is the linear change per degree of global warming for the specific variable and grid cell. Spatial fields of these coefficients, across the whole domain of interest, constitute the ''pattern'' in PS. A normalized (i.e., local change per degree of global warming) pattern is diagnosed from one or more simulations with a GCM.

The best way to diagnose the patterns is an important consideration, with two problems to address. First, the response to an external forcing can be obscured by internal climate variability, causing the diagnosed pattern to differ from the true response of the model. This problem can be reduced in three ways (Mitchell 2003): initial-condition ensembles (where available) can be averaged to strengthen the signal-to-noise ratio; patterns can be diagnosed by regression over time (with appropriate time filtering) instead of simply differencing two periods; and patterns can be diagnosed simultaneously from several runs of the same GCM under different forcing scenarios (e.g., RCP2.6 and RCP4.5 data pooled together) rather than diagnosed from a run of the GCM under a single climate change scenario. By adopting all three strategies (Osborn et al. 2016), ClimGen is likely to represent the model's response to climate change forcing more accurately (see section SM1 in the supplemental material for further details). However, pooling data over time and from all RCP simulations can exacerbate the second problem, namely that nonlinear GCM behavior or differences in regional forcing may be manifested by differences in patterns between scenarios or over time within one scenario, thus violating the linear assumption of Eq. (1). In this study,

TABLE 1. The various combinations of RCP-forced transient GCM data used to generate the pattern coefficients needed for the PS projections.

Pattern name	HadGEM2-ES simulations (1951–2100) used to diagnose pattern
RCPall	RCP2.6, RCP4.5, RCP6.0, and RCP8.5
RCP264560	RCP2.6, RCP4.5, and RCP6.0
RCP26	RCP2.6 only
RCP85	RCP8.5 only

therefore, we also diagnose different patterns from subsets of the RCP simulations to assess the impact on PS performance.

Within ClimGen, separate fields of α are diagnosed for each of the 12 calendar months and used in Eq. (1) to produce climate projections at the monthly time scale inclusive of changes in the annual cycle. We focus here only on changes to mean near-surface temperature over land (where PS is most commonly applied) and do not consider changes in variability or over the oceans. For the purpose of this paper, we diagnose the patterns and apply PS on the GCM's native grid (whereas ClimGen is applied after interpolation to a 0.5° latitude × 0.5° longitude land-only grid) so that we can validate the PS climate projections against the actual GCM transient climate data.

The normalized change patterns (i.e., the fields of α coefficients) required for PS are diagnosed from HadGEM2-ES data for the 1951–2100 period using all available ensemble members [see section SM1 and Osborn et al. (2016) for details]. Alternative patterns are calculated using data pooled from different RCP permutations (Table 1). Some permutations purposely exclude the RCP8.5 GCM data from the pool to enable the validation of PS projections (using RCP8.5 data in this case) to be independent of the data used to diagnose the pattern. Exploring the sensitivity of PS errors to the data used to diagnose the fields of α coefficients is an important aspect of the validation exercise, since PS is often applied using patterns calculated from one RCP simulation to create a projection under ΔT_t from another scenario, with no consideration of the pattern dependence of the GCM under each scenario. Once the patterns have been diagnosed, PS projections are then calculated by combining each of the pattern permutations in Table 1 with ΔT_t from Fig. 1.

b. *GCM data for evaluating the performance of PS projections*

Throughout this analysis we use PS to attempt to emulate the transient near-surface air temperature response of the HadGEM2-ES climate model, a coupled ocean–atmosphere circulation model with dynamic vegetation, land and ocean carbon, and tropospheric chemistry components (Collins et al. 2011). Since a focus of this study is the performance of the PS method under high-end warming scenarios, we use HadGEM2-ES climate data from the CMIP5 RCP8.5 scenario as the validating data (i.e., the target climate to be attained by the PS projections). The RCP8.5 HadGEM2-ES data available are as follows:

(i) a four-member ensemble covering 2001–2100, with greenhouse gas forcing not yet stabilized (Jones et al. 2011), and
(ii) a single-member extension to 2299, with CO_2 concentration stabilizing by 2250 at 2000 ppm (Caesar et al. 2013).

These simulations are appended to an ensemble of runs under historical forcing and four-member ensembles for each of the other three RCPs (RCP2.6, RCP4.5, and RCP6.0) are also used for diagnosing normalized change patterns. For making the PS projections under RCP8.5 [via Eq. (1)], the driving ΔT_t is the global-mean annual air temperature simulated by HadGEM2-ES for 1951–2299 derived from the historical simulations followed by both parts of the RCP8.5 data described above. Ensemble means are used to 2100 and the single run thereafter. The ΔT_t time series is filtered with smoothing splines to isolate the forced climate change signal from unforced interannual variability and is expressed as anomalies from the simulated 1961–90 mean. Using 1961–90 as the reference baseline means that the resulting PS projection data represent changes (anomalies) from this period also. For validation of the resulting PS projections, the individual gridcell HadGEM2-ES

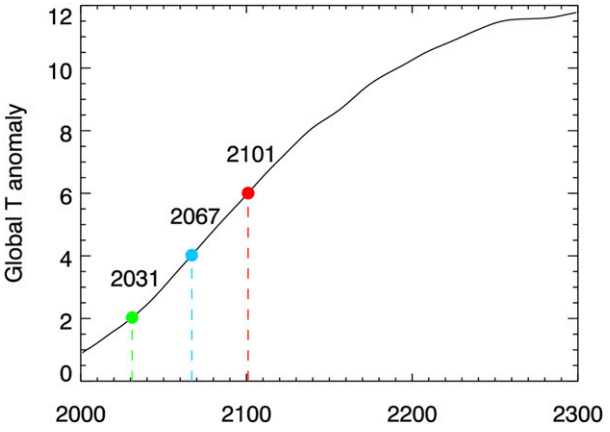

FIG. 1. Smoothed HadGEM2-ES ΔT_t (RCP8.5 monthly) used to make the PS projections (shown here relative to the 1861–90 mean). Green, blue, and red locations show years where SWLs of 2°, 4°, and 6°C, relative to 1861–90, are reached (2031, 2067, and 2101, respectively). Using smoothed data increases the likelihood that reaching the SWL arises from the climate change signal, rather than realization-dependent unforced variability.

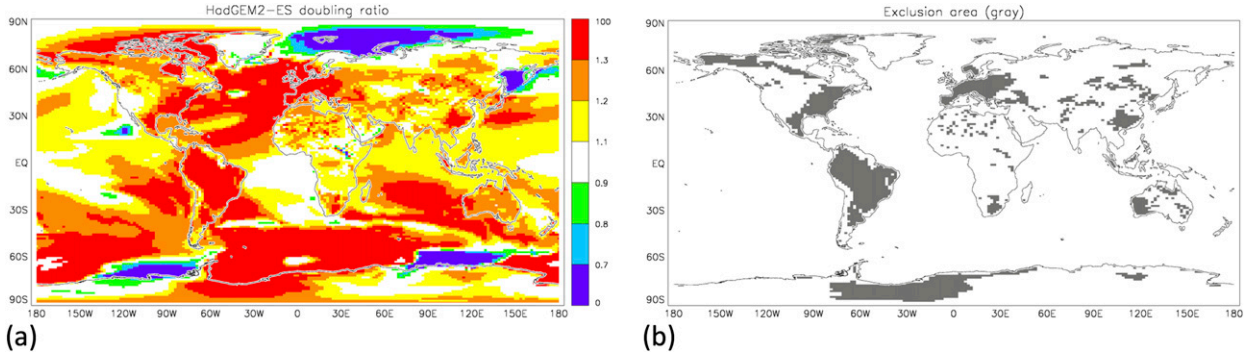

FIG. 2. (a) HadGEM2-ES doubling ratio [after Good et al. (2015)]. (b) Land grid cells where the doubling ratio is less than 0.75 or greater than 1.25 (gray shading). These cells are excluded to leave the land area where the HadGEM2-ES climate response to greenhouse gas forcing is approximately linear.

GCM data are also anomalized to the 1961–90 period and a running 30-yr mean applied to isolate the climate change signal.

As well as validating PS performance over the course of the RCP8.5 simulation, we also look specifically at years when global warming reaches SWLs of 2°, 4°, and 6°C above preindustrial values to see if PS performance is compromised at high-end warming levels. To identify these years, we add 0.35°C [the difference in global-mean temperature between the means of the 1961–90 and 1861–90 periods in the observational HadCRUT4 dataset of Morice et al. (2012)] to the ΔT_t time series. The 1861–90 period is used to approximate the preindustrial level because it is the earliest 30-yr period with land and marine instrumental data in both hemispheres. The adjusted annual ΔT_t series (Fig. 1) is then used to select illustrative SWL years (the 30-yr means centered on these years are used; see section SM1.2). Note that PS projection errors during these periods are indicative of PS performance for these SWLs, although strictly only when they are reached by following the RCP8.5 scenario because the GCM-simulated patterns could differ if, say, the SWL of 2°C was reached under a more slowly increasing scenario (we partly address this by comparing patterns diagnosed from different scenarios).

The gridded HadGEM2-ES validation data can be used as either an ensemble mean or as single realizations. We use both to investigate the contribution of unforced variability to the assessment of PS performance, but the ensemble mean is limited to pre-2100 and thus excludes the 6°C SWL (Fig. 1) centered on 2101 and thus requiring the 2086–2115 30-yr mean.

c. *Assessing the influence of climate system nonlinearities*

Errors in PS projections arise from errors in the diagnosed patterns and from nonstationary patterns for some climate variables between scenarios and over time within a scenario. Nonstationary patterns, violating the PS assumption of linearity between local change and global temperature change, may arise through different regional forcings or through a nonlinear response of the simulated climate system. We attempt to isolate the component from nonlinear climate system behavior by using prior identification of nonlinear behavior in HadGEM2-ES (Good et al. 2015). Good et al. (2015) compared changes in air temperature after two successive CO_2 doublings to derive a local (i.e., grid cell) linearity metric, the doubling ratio:

$$\frac{\Delta V_{\mathrm{db2}}}{\Delta V_{\mathrm{db1}}}, \tag{2}$$

where ΔV_{db1} is the change in gridcell surface air temperature to a $2 \times CO_2$ state, and ΔV_{db2} is the further change to a $4 \times CO_2$ state. The pattern of the doubling ratio (Fig. 2a; after Good et al. 2015) indicates regions where warming after the second doubling is greater than after the first (red) and where it is less (blue). Good et al. (2015) attribute the red Atlantic and European sectors to the nonlinear weakening of the Atlantic meridional overturning circulation (AMOC) affecting regional air temperature. This behavior will be GCM-dependent, though Sgubin et al. (2015) find that HadGEM2-ES AMOC changes are similar to other CMIP5 GCMs during transiently increasing forcing, and that model dependence arises principally after a switch to decreasing forcing. High-latitude features with doubling ratios below one (blue) are related to changes in snow or sea ice that are rapid under the first doubling but then stabilize (thus weakening feedbacks) in the second doubling. Zones of high doubling ratios over South America arise from a combination of vegetation, precipitation, and soil moisture dynamics influencing the ratio of sensible to latent heat fluxes (Good et al. 2015); this ratio changes more strongly in HadGEM2-ES than

in the other four GCMs they examined, so it is likely that this nonlinearity is less pronounced for other GCMs than HadGEM2-ES.

To investigate the influence of these known nonlinearities upon the PS projection errors, we define a spatial mask (Fig. 2b) to exclude specific grid cells from the error metrics presented in section 3c, excluding all cells where the doubling ratio is <0.75 or >1.25 (note that we already exclude ocean areas from our analysis). Although the threshold choice is arbitrary, it provides one benchmark for identifying the influence of nonlinearities on PS performance.

3. Results

a. Validation of pattern-scaling performance

Metrics comparing each PS projection against the transient GCM (HadGEM2-ES) data illustrate the ability of the ClimGen pattern scaler to capture the behavior of the GCM. We first consider performance metrics where the GCM data (the ''target'' data that we attempt to reproduce using PS projections) are from the single-member HadGEM2-ES simulation covering the full 2001–2299 period. This enables us to explore performance at very high warming levels out to 2299, but the single realization prevents us from quantifying the effects of unforced climate variability (examined in section 3b using all four ensemble members available to 2100). Validation results considering climate system nonlinearity are described in section 3c.

Patterns of differences between the PS projection and the single GCM ensemble member for representative 2°, 4°, and 6°C SWL periods are shown in Fig. 3, as well as PS–GCM differences using the ensemble mean for 2° and 4°C. In these comparisons the PS projections are generated using the so-called RCPall pool of training data, since this is the ClimGen default and has been the basis for constructing climate scenarios for the impact work referenced earlier. Broadly, there are differences from the single GCM run of both signs and with magnitudes that are mostly between 0° and 1.25°C for SWLs of 2° and 4°C. In some regions (e.g., eastern North America), the PS–GCM differences have opposite signs at 2° and 6°C, perhaps indicating a nonlinear warming pattern with the GCM data lying above the linear PS regression line during one period and below it during another.

However, many of these PS–GCM differences are much smaller when the GCM ensemble mean is used (only possible for SWLs of 2° and 4°C), with absolute differences nearly all less than 0.75°C and many less than 0.25°C even for SWL of 4°C. There is a tendency for the PS projection to be biased cool over land, especially for the 4°C SWL, although this is not ubiquitous (e.g., there is a warm bias over Asia at 2°C SWL in January). For the higher SWL of 6°C (year 2101 in HadGEM2-ES RCP8.5) stronger biases appear, with PS underestimating the GCM warming in the Amazon and around the Arctic but overestimating the GCM warming over the land around the North Atlantic in January.

These PS–GCM differences are large enough to have practical significance [see the regional damage functions of Arnell et al. (2018) for examples of the regional impacts arising from differences in temperature change] but are nevertheless small compared both to the climate change signal and to the differences between GCM projections (Heinke et al. 2013). Tebaldi and Arblaster (2014) confirmed the overall validity of the pattern-scaling approach as an approximate representation of the CMIP3 and CMIP5 multimodel ensembles for land air temperatures. This ''validity'' arises because errors in pattern scaling were shown to be small relative to the ensemble intermodel spread [Fig. 4 of Tebaldi and Arblaster (2014) shows that the spread of change patterns from different models is much larger than the spread of change patterns from one model under different RCP scenarios, which represents nonstationary or nonlinear behavior that PS cannot always capture]. Osborn et al. (2016) quantified similar results for the CMIP5 ensemble: for annual temperature, around 10% of the local variance across the ensemble arises from differences in the normalized patterns of change between scenarios for the same model. There is some spatial variation, but in only a few locations does the contribution rise above 20%.

Here, we add detail to these results by calculating the time-evolving magnitude of the pattern-scaling ''error''[1] compared to the magnitude of the GCM projection itself. The root-mean-square (RMS) of all land gridcell differences is much smaller than the RMS of the (30-yr running mean) GCM land temperature change fields that the PS projection is attempting to reproduce. Their ratio decreases from about 0.3 to about 0.1 as global warming approaches 3°–4°C, after which the ratio gradually rises. This metric aggregates over all land grid cells, including some where the local PS–GCM differences may nevertheless be quite large (Fig. 3); the local error results are considered in more detail later. The

[1] Note that ''error'' is in quotation marks when we refer to the difference between a PS projection and the corresponding GCM projection, because this difference can arise through internal variability in the GCM simulation as well as through errors in pattern scaling.

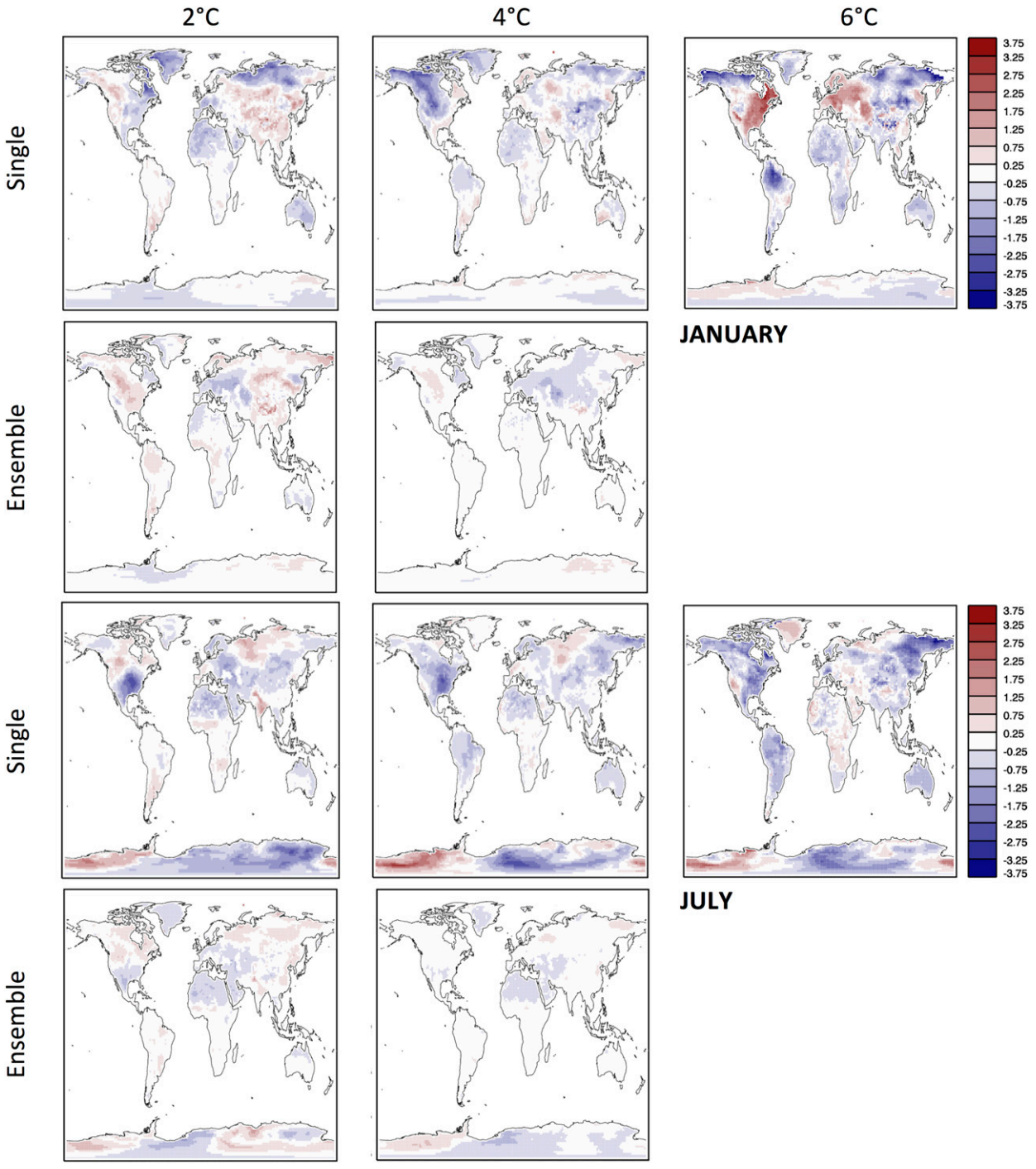

FIG. 3. Differences (PS − GCM) between PS and (first and third rows) single-member GCM projections or (second and fourth rows) ensemble-mean GCM projections of land air temperature change (°C) for (top two rows) January and (bottom two rows) July for periods when ΔT_t = (left) 2°, (center) 4°, and (right) 6°C under RCP8.5. PS projections are generated using the RCPall pattern.

relative size of the PS "error" follows this pattern consistently for all months of the year examined here (Fig. 4, individual lines). This suggests that PS is able to represent the Arctic amplification of warming in winter months equally as well as the more moderate warming projected in other months, even though the amplification is associated with nonlinear snow–albedo feedback over high-latitude land.

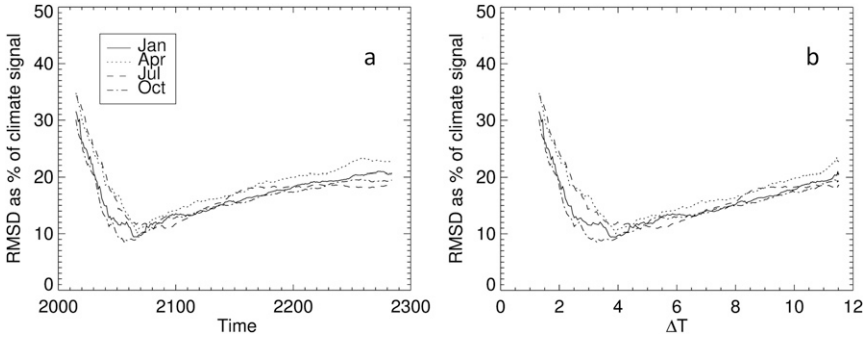

FIG. 4. Global land RMS difference (°C) between the gridded PS projection (using the RCPall pattern) and the single-member GCM projection of land air temperature change under RCP8.5. The RMS difference is expressed as a percent of the GCM climate change signal itself (i.e., the global land RMS of the GCM gridded temperature change relative to simulated 1961–90) and is plotted as a function of (a) time and (b) the GCM global ΔT_t.

This comparison is not wholly independent because HadGEM2-ES RCP8.5 data are used in both the derivation of the RCPall pattern and in the testing of the PS projection based on this pattern. The comparison beyond 2100 (or ~5.5°C of global warming) *is* independent, since no GCM data beyond 2100 were used in the pattern diagnosis. Completely independent tests can be made of PS projections using a pattern diagnosed from all simulations except RCP8.5 (i.e., the RCP264560 pattern) and show very similar results (cf. the green and black lines in Fig. 5), so this lack of complete independence is not giving an overly optimistic view of PS performance.

The RMS differences (Fig. 5) for the remaining pattern permutations considered (Table 1) show comparable levels of performance for lower global warming levels, for which there is a low level of emission dependence. At higher warming levels, however, the RCP85 pattern is superior. The RCP26 pattern performs least well; we might attribute this, partly, to a less well-defined pattern of coefficients from the RCP2.6 ensemble because of its weaker forcing, which is then extrapolated to emulate high-forcing responses with characteristics not present in the RCP2.6 training data. The differences between most PS projections commence at approximately the ΔT_t = 3°–4°C (2050–70 for HadGEM2-ES RCP8.5), which must be linked to nonstationarity of the patterns because of, for example, higher dependence on emission scenario.

Similar results are obtained for five more CMIP5 GCMs (see section SM2.1): the RCPall pattern performs slightly better than the RCP85 pattern for specific warming levels up to approximately 3.5°C above preindustrial for CanESM2 and up to approximately 3.0°C for CCSM4, CSIRO Mk3.6.0, and IPSL-CM5A-LR (acronym expansions are available online at http://www.ametsoc.org/PubsAcronymList). For CNRM-CM5, which warms the least under RCP8.5 out of the six GCMs analyzed in this study, the RMS difference between the GCM and the PS projections is generally larger and shows an earlier divergence between the patterns such that PS with the RCP85 pattern has a smaller error than with RCPall once global warming exceeds about 2°C.

The spatial patterns of ''errors'' for the best (RCP85) and worst (RCP26) performing PS projections near to ΔT_t = 3°–4°C illustrate the geographical source of their performance disparities (Figs. 6a,b,d,e) compared to the GCM climate change signal (Figs. 6c,f) for the 4°C SWL. Both patterns tend to underestimate the HadGEM2-ES warming over land overall, but the regional differences (of both signs in winter) are clearly stronger in the RCP26 pattern for both the Northern Hemisphere (NH) winter (Fig. 6a) and summer (Fig. 6d). It is possible that this arises because patterns evolve differently over time between RCP scenarios (nonlinear dependence on forcing strength or regional differences in forcing) so that the RCP26 pattern is simply not able to emulate the RCP8.5 scenario very well. Alternatively, the local climate response may be linear, but that the change pattern is more dominated by sampling variability when it is diagnosed from just the RCP2.6 scenario, with a tendency to underestimate the slope of the local to global relationship and thus for PS to underestimate the projected warming.

It is also useful to examine PS projection ''errors'' as a function of *local* (i.e., grid cell) warming (hereinafter local ΔT) as opposed to global mean warming ΔT_t. Examining local errors against local ΔT can tell us more about the conditions under which PS performs well or poorly. We define local ΔT as the surface air temperature change in each grid cell, relative to 1861–90, in the

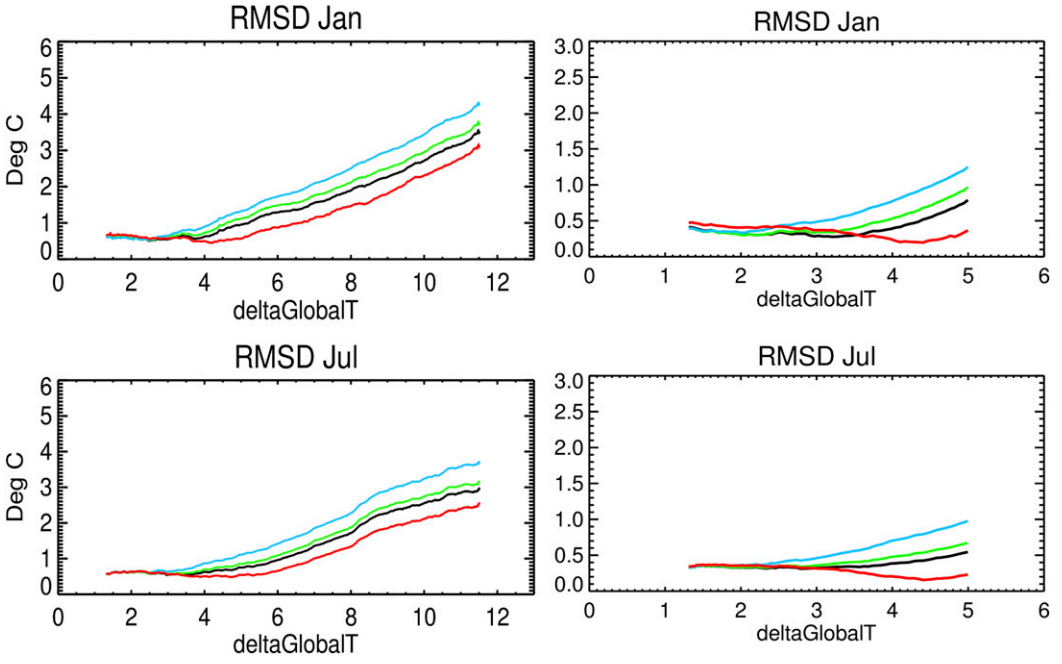

FIG. 5. Global land RMS difference (°C) as a function of ΔT_t, between gridded PS projections [using patterns RCPall (black), RCP264560 (green), RCP26 (blue), and RCP85 (red)] and (left) the single-member GCM projection (2001–2299) under RCP8.5 and (right) the ensemble-mean GCM projection (2001–2100). Results are shown separately for (top) January and (bottom) July. Note the different axis ranges for the left and right columns.

validation GCM dataset and plot PS projection errors in each cell as a function of local ΔT for the global 4°C SWL (Fig. 7). The final height of curves in Fig. 7 reflect total accumulated errors for each PS projection while curve steepness relates to the accumulation of errors at a specific local ΔT. Error accumulation can be attributed to either a poor PS performance or a greater density of gridcell counts at that local ΔT. Quintile positions of the gridcell local ΔT populations of each given month are shown in Fig. 7 to show the distribution of cell counts according to their local ΔT values.

For January (Fig. 7a) the performance of RCPall is comparable to RCP85 through most local ΔT values, only diverging for grid cells with local warming of more than 10°C, while the accumulated errors for other pattern permutations (especially RCP26) grow at lower local ΔT values. The divergence of the RCPall and RCP85 error curves in Fig. 7a corroborates the correlation of the January spatial errors shown in Fig. 6a with geographical locations of stronger warming shown in Fig. 6c.

For July (Fig. 7b) the range of local ΔT is narrower than January (attributable to less summer warming in the Northern Hemisphere), but the overall PS performance rankings are the same as January, as are the approximate total accumulated error values. At first glance, the similarity of the July and January accumulated errors is surprising if we assume that PS performance degrades with rising local ΔT, since the GCM July warming is much lower than January (cf. Figs. 6f and 6c and the range of local ΔT in Fig. 7). However, the spatial patterns of July errors (Figs. 6d,e) show that strong PS errors exist in July over Antarctica, for both RCP26 and RCP85 patterns, despite only moderate local ΔT (Fig. 6f). Note that PS is not typically applied over Antarctica and the standard version of ClimGen does not include Antarctica because of insufficient observational data to combine with the PS projections.

b. *Quantifying GCM internal variability as a source of PS–GCM projection differences*

Comparing PS projections to a single GCM simulation is useful but it is not a perfect measure of PS performance at emulating the GCM "climate change signal," since individual GCM realizations have a unique, internally driven climate component, independent of the externally forced climate change signal. RMS differences considered so far, therefore, are a combination of any deficiencies in PS emulation of the GCM externally forced climate change signal *and* the unforced variability simulated by the GCM on time scales of 30 yr or longer (unforced variability on shorter time scales will not inflate RMS differences because we compare 30-yr running means). This was already visible in the much

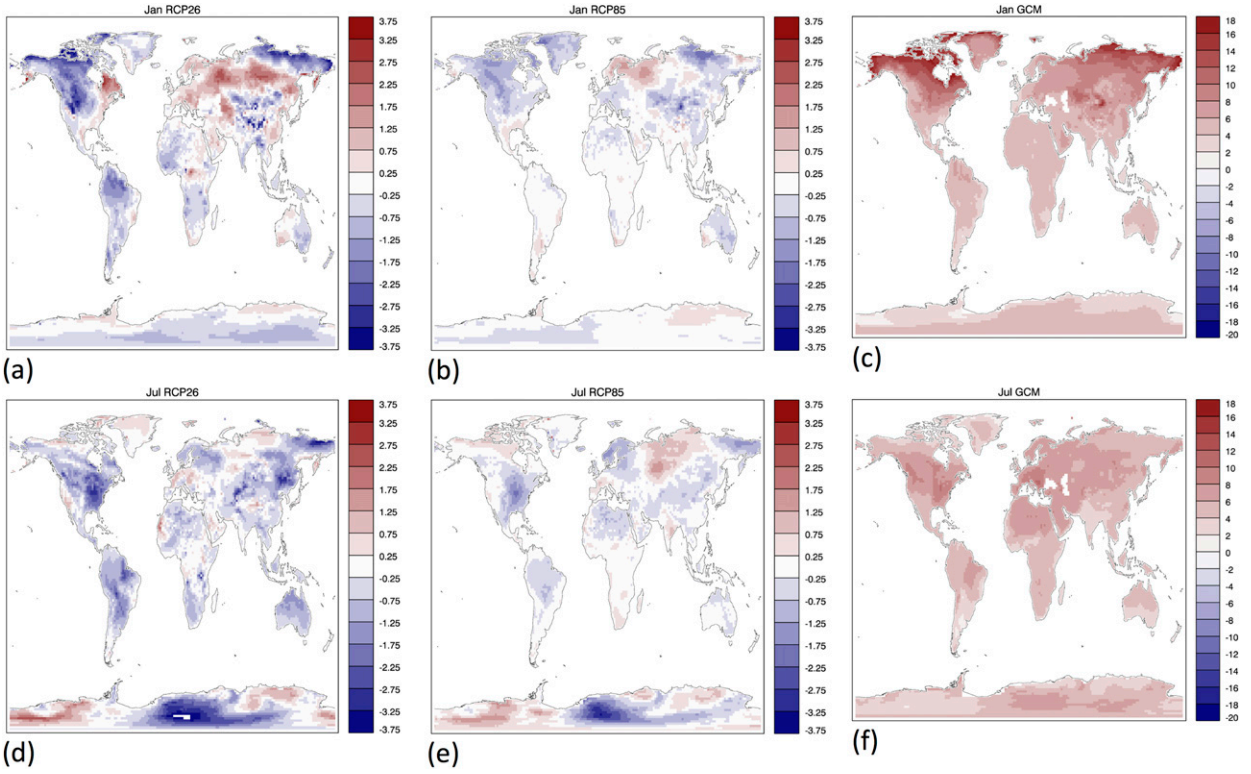

FIG. 6. Differences (°C) between PS and single-member GCM projections of (a),(b) January and (d),(e) July land air temperature change for the period when $\Delta T_t = 4$°C under RCP8.5. PS projections are generated using the patterns RCP26 in (a),(d) and RCP85 in (b),(e). The GCM projected climate change anomalies (°C) for (c) January and (f) July are also shown.

weaker PS–GCM difference patterns when comparing with the ensemble mean than with a single GCM run (Fig. 3). We can quantify these contributions to the overall PS–GCM differences by recalculating global land RMS differences between the PS projection and the mean of all four HadGEM2-ES RCP8.5 ensemble members, limited to the 2001–2100 period when all four ensemble members are available.

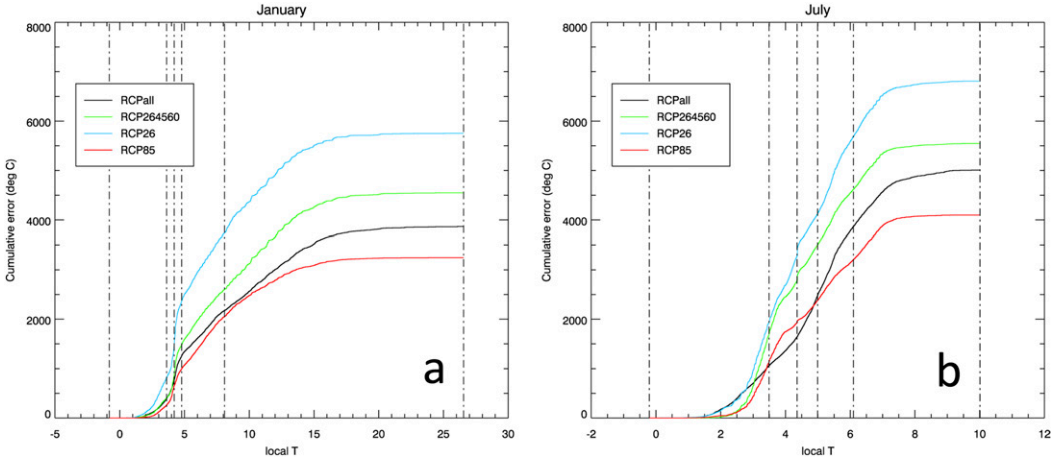

FIG. 7. Cumulative land gridcell absolute differences (y axis, °C) between PS projections [RCPall (black), RCP264560 (green), RCP26 (blue), and RCP85 (red)] and the GCM projection as a function of increasing gridcell local ΔT projected by the GCM (x axis, °C). Data are for (a) January and (b) July for the period when global ΔT_t = 4°C under RCP8.5. Vertical lines indicate quintiles of gridcell local ΔT projected by the GCM (i.e., equal counts of land grid cells lie between each pair of vertical lines).

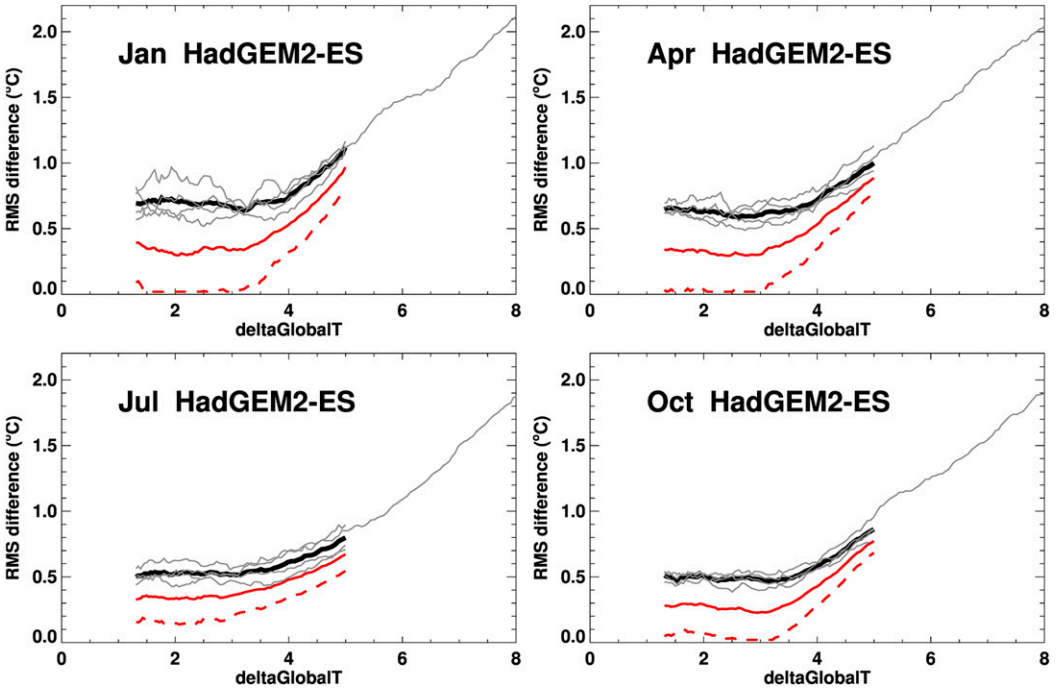

FIG. 8. Global land RMS differences (°C) between gridded PS projections (RCPall) and each single-member GCM RCP8.5 projection (gray; only one member extends to 2299) and the ensemble-mean GCM projection (red; ends in 2100 when three of the four ensemble members stop), under RCP8.5, as a function of ΔT_t. The mean of the individual ensemble member results is shown in black, and the red dashed line indicates the inferred RMS difference from a hypothetical infinite ensemble. Results shown separately for January, April, July, and October.

RMS differences show marked reductions when comparing with the ensemble mean rather than with a single GCM simulation (Fig. 8 for PS projection RCPall). Given that the unforced variability in each ensemble member is largely independent [see section 9.1 of Jones et al. (2011)], the standard deviation of unforced variability at each grid cell in the ensemble mean will be reduced by the square root of the ensemble size. Thus with a four-member ensemble, we expect the component of the PS–GCM difference arising from GCM internal variability to halve compared with a single ensemble member, while the difference arising from the genuine error between the PS projection and the GCM response to the RCP forcing will be unchanged.

For global warming up to about 3°C, RMS differences are approximately halved in all months except NH summer (red line compared with the black mean of the individual gray lines in Fig. 8), suggesting that the genuine PS projection error is very small. Beyond ΔT_t = 3°C, the externally forced climate signal becomes even stronger compared with the internal variability and so the reduction in PS–GCM differences diminishes, visible in a steepening of the ensemble-mean RMS curve toward higher levels of warming (Fig. 8, red).

If we had an infinite ensemble of RCP8.5 runs from HadGEM2-ES, we would expect the RMS difference arising from the unforced GCM variability to be reduced by the same amount as the reduction already seen in going from one to four ensemble members. In other words, the four-member ensemble mean still contains significant levels of unforced variability, which unfairly penalizes the apparent performance of PS at emulating the GCM. Doubling the reduction in RMS differences from the mean of the single-run results (Fig. 8, black lines) to the four-member results (red lines) gives an estimate of the RMS differences with a hypothetical infinite ensemble with no unforced variability (Fig. 8, red dashed lines). This indicates that the genuine RMS difference between PS and GCM projections is close to zero (except in NH summer) for global warming up to 3°C under RCP8.5, for HadGEM2-ES. The performance then deteriorates significantly for increased warming, although a fair evaluation of PS performance would still show smaller errors than the single or four-member results. Even for the month with the largest residual error (July), the inferred PS RMS error is less than 0.25°C for global warming up to 3.5°C. Note the earlier caveat that the RCP8.5 testing data were also partly used to define the RCPall pattern used to make these PS projections.

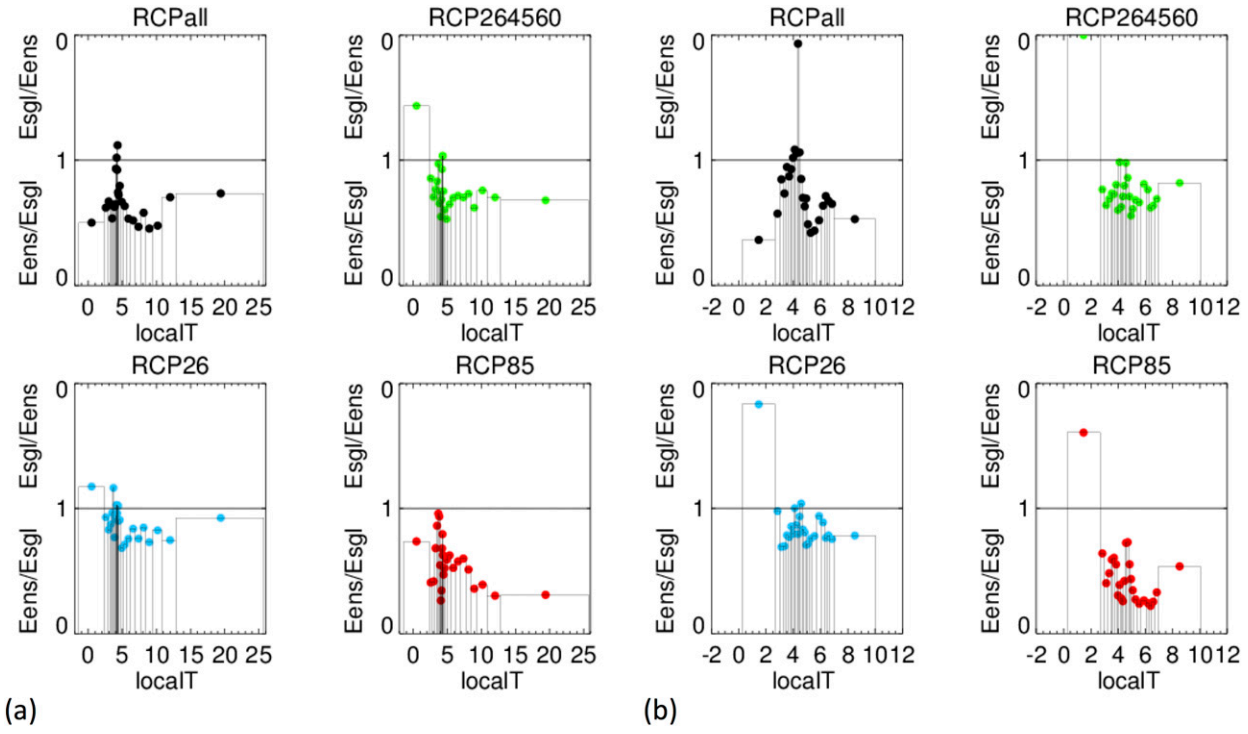

FIG. 9. The ratio of absolute differences between gridcell PS projections and either single-member (Esgl) or ensemble-mean (Eens) GCM projections as a function of gridcell local ΔT projected by the ensemble-mean GCM (x axis, °C) for the period when $\Delta T_t = 4$°C under RCP8.5. Each panel is for a different PS projection, RCPall (black), RCP264560 (green), RCP26 (blue), and RCP85 (red), for (a) January and (b) July. The PS–GCM absolute differences were first averaged over local ΔT bins containing equal numbers of grid cells (indicated by the gray histograms, which therefore also indicate the density distribution of gridcell temperature changes). Where Eens < EsgI, Eens/Esgl is plotted in the lower half of the panel; where Eens > EsgI, Esgl/Eens is plotted in the upper half of the panel and the y-axis range is reversed.

Similar results are obtained for the five other CMIP5 GCMs analyzed (see section SM2.2): inferred RMS errors are also less than 0.25°C for global warming up to at least 4.0°C for CanESM2, CCSM4, and IPSL-CM5A-LR, up to 2.1°–3.4°C (depending on month) for CNRM-CM5, and up to 3.4°–4.2°C for CSIRO Mk3.6.0. For all GCMs except CNRM-CM5, PS performance is more favorable than found for HadGEM2-ES.

We also investigate how this apparent reduction in PS error (when comparing with a GCM ensemble mean) varies as a function of gridcell climate change, local ΔT, as per Fig. 7, but using local ΔT from the ensemble-mean GCM data. To reduce the noise that would result if the PS–GCM differences were plotted for each of the 9244 land grid cells, the grid cells are first grouped into 25 bins. Each bin is defined to contain all grid cells with a particular range of GCM-simulated local ΔT, with the ranges chosen so that the grid cells are divided equally into the 25 bins (so each contains 369 grid cells, apart from the final bin). We use local ΔT from the GCM ensemble mean even when considering the PS–GCM difference for the single GCM ensemble member, because the ensemble mean is a truer representation of the climate change that the PS projection is attempting to emulate and it also ensures that the same grid cells are assigned to each bin regardless of the comparison being made. The PS–GCM projection absolute differences at each grid cell are averaged over the bin to obtain the ratio Eens/Esgl, where Eens and Esgl are the bin-averaged PS–GCM differences using the ensemble mean and a single ensemble member, respectively. In Fig. 9, we plot the Eens/Esgl ratios for the 4°C SWL period against the bin-averaged local ΔT and the width of the bins indicates the density of local ΔT values (because each bin contains the same number of grid cells).

Where Eens/Esgl < 1, the PS–GCM difference is reduced by comparing with the GCM ensemble mean rather than with a single GCM run for the 4°C SWL. As expected, this occurs across most local ΔT bins for both January (Fig. 9a) and July (Fig. 9b). Improvements are largest for the RCPall and RCP85 PS projections and less clear when scaling the RCP26 pattern, suggesting a higher contribution in the latter from actual pattern deficiencies (in terms of capturing the forced climate signal). For the RCP85 pattern the improvements are greatest over regions with higher local warming,

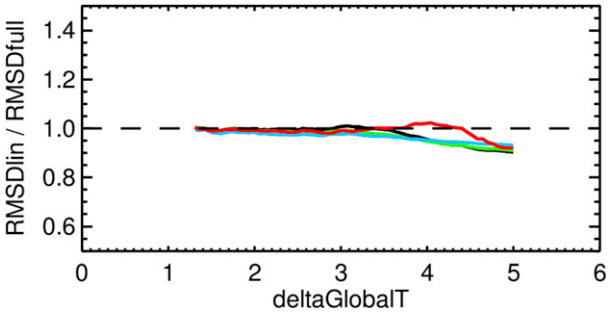

FIG. 10. The ratio of the RMS differences between PS projections [RCPall (black), RCP264560 (green), RCP26 (blue), and RCP85 (red)] and the ensemble-mean GCM projection of annual-mean temperature change when they are calculated over a limited set of land grid cells (RMSDlin, where the response of HadGEM2-ES to CO_2 forcing is approximately linear; see Fig. 2b) compared to when the differences are calculated over all land grid cells (RMSDfull), as a function of global warming (°C).

especially in January (Fig. 9a). For a very few bins, notably those with least local warming, PS performance actually deteriorates when comparing with the GCM ensemble mean (i.e., Eens > Esgl, and the ratio is plotted as Esgl/Eens on the inverted scale in the top half of each panel). The reduced PS–GCM projection errors are not homogenous across all bins. When using the RCPall pattern, there are four bins close to the center of the distribution (with local ΔT close to 4°C, which is also a very dense part of the distribution in January; Fig. 9a) with either no improvement or even deterioration when PS is compared to the ensemble mean. For other bins, for instance those from local ΔT between 5° and 10°C, the use of the GCM ensemble mean approximately halves the PS–GCM differences.

c. *Quantifying GCM nonlinearity as a source of PS error*

Genuinely poor PS performance would arise from two main sources: 1) nonlinearity in the response to the same type of forcing, with the response pattern varying over time or as the forcing strengthens, and 2) differences in the response patterns between scenarios because of different regional forcings (especially aerosol and land-use changes). To assess the contribution from the first of these sources, we apply a spatial mask to the evaluation of PS projection performance as described in section 2c. This mask excludes grid cells where separate simulations (Good et al. 2015) have already demonstrated that HadGEM2-ES responds nonlinearly to increased CO_2 forcing. Application of the mask excludes 2469 grid cells (27% of the land grid cells). We apply the mask to the PS–GCM projection differences, using the HadGEM2-ES ensemble mean to reduce the contribution of internal variability to the difference. The comparison is made using mean annual temperature changes because annual means were used to generate the mask from Good et al. (2015), and we consider PS projections from the same four pattern permutations (Table 1).

Figure 10 compares the RMS of the PS–GCM differences between the masked (RMSDlin) and unmasked (RMSDfull) global land fields over the course of the RCP8.5 simulation. Masking known regions of nonlinearity (for this GCM, HadGEM2-ES) does decrease the PS–GCM differences but by no more than 10% and not until global warming reaches 4°C. Furthermore, the PS projection made using the RCP85 pattern (red line, Fig. 10) shows no decrease until later in the simulation. These results suggest that nonlinear responses within HadGEM2-ES make only a small contribution to the PS–GCM land temperature differences for global warming levels up to 5°C at least.

Analyzing the pattern of PS–GCM differences using a binned approach similar to Fig. 9 (not shown) for a global SWL of close to 5°C where the improvement is greatest (Fig. 10), shows that exclusion of grid cells with nonlinear behavior reduces the PS–GCM differences particularly in regions with local warming around 6°C or with very high levels of local warming from approximately 10°–18°C.

4. Discussion and summary

We have investigated the performance of the popular PS technique, as implemented by the ClimGen pattern scaler (Osborn et al. 2016), for emulating the climate change response of the HadGEM2-ES under the high-end emission scenario RCP8.5. We repeated the analysis with five other CMIP5 GCMs (see section SM2) and obtained similar key findings. We focused on climate changes over land, where the linear assumptions that underlie PS are more reasonable and where PS is most often applied. We evaluated changes in near-surface air temperature, though the approach is applicable to other variables (a companion paper will report our findings for precipitation; C. J. Wallace et al. 2018, unpublished manuscript). We have paid particular attention to the impact on the performance metrics of the unforced climate variability present in the GCM simulations and of nonlinearities in the climate change response of the HadGEM2-ES.

Initially we show that even when unforced variability is not accounted for, the difference between the PS and GCM projections when evaluated over the global land surface is only 10%–15% of the GCM climate change response itself for a wide range of global warming (2°–7°C; Fig. 4). This is much smaller than other sources of

uncertainty in climate change projections such as the choice of forcing scenario and the spread among an ensemble of multiple climate models. It is consistent with other reported assessments [e.g., Table 2 of Heinke et al. (2013), which assesses the linear change signal in CMIP3 GCMs].

Using the mean of a four-member initial-condition ensemble of GCM runs reduces the standard deviation of unforced variability by half, and the difference between the PS projections and the direct GCM results is then notably reduced (Fig. 8). Indeed, for PS projections using patterns diagnosed from a range of different GCM simulations, the PS–GCM difference is almost halved for a range of global warming from 1° to 3°C, indicating that the majority of the remaining PS–GCM difference can be attributed to the residual internal variability in the four-member GCM ensemble mean. This cannot be confirmed since we do not have an infinite GCM ensemble and the inference that there is almost zero error in the PS approximation for this range of global warming range should be considered cautiously. PS cannot perfectly represent even the linear component of the forced climate response because there will be some contamination of the training data (used to diagnose the patterns) by unforced variability itself, although we mitigate this by pooling multiple temporally smoothed simulations across all RCPs and regress over the entire 1950–2100 period to generate the RCPall pattern. Lynch et al. (2017) also find reduced bias and mean errors when patterns are diagnosed using linear regression compared with the "delta" method.

We recommend, therefore, that PS performance should not be evaluated by comparison against a single GCM simulation (or even a small ensemble) without carefully considering the role of unforced internal GCM variability as a cause of the differences found. Heinke et al. (2013) compare the overall variance of the residuals against the unforced variability simulated in the GCM control runs to address this issue. When evaluated against the HadGEM2-ES climate change signal rather than a combination of signal and unforced variability, the errors arising from the PS technique are very small (inferred root-mean-square errors of 0.25°C or less when land monthly temperature changes are aggregated globally) for global warming up to 3.5°C. For global warming greater than this, the PS projection errors grow mostly as a result of scenario dependence in the GCM results but with a contribution from nonlinearity in the GCM response. The climate change signal strengthens too, so the error remains small relative to the climate signal.

We also evaluated the performance of patterns diagnosed from different sets of GCM simulations, and summarize the evaluation against the four-member ensemble mean in Fig. 5 (cf. right and left columns, which shows the performance against the single run that extends to much greater levels of warming). The PS performance using patterns diagnosed from all RCP runs pooled together [RCPall (black line), the default in ClimGen] is similar to the RCP264560 (green) pattern that is completely independent of the validation GCM data (RCP8.5). They perform best for global warming up to approximately 3.5°C above preindustrial, but beyond this the pattern diagnosed from only the RCP8.5 data (RCP85, red) clearly performs better. Although this is not an independent test (using the same simulations to diagnose the patterns and to serve as the validation dataset is likely to overestimate performance), we nevertheless recommend using patterns diagnosed from strong forcing scenarios when making PS projections under high-end global warming scenarios. For some months, the RCP85 pattern performs poorly under global warming of 2°C. The pattern diagnosed using only RCP2.6 simulations (RCP26, blue) provides no advantages over the RCPall pattern even for small amounts of warming early in the RCP8.5 projection, and its performance deteriorates earlier.

Since RCPall is the default in ClimGen configuration used to generate climate scenarios for impact work, these results suggest an improvement could be made by using RCPall for warming up to approximately 3.5°C and then the RCP85 pattern for projections of high-end warming. A transition period might be used to avoid discontinuities. This recommendation applies when emulating other GCMs too (see section SM2.1).

We examined the pattern of PS–GCM projection differences geographically and as a function of local warming. There is only limited correspondence between the largest differences and regions that had previously been identified as having a nonlinear response to CO_2 forcing in this GCM (Good et al. 2015), underlining the small contribution of nonlinear climate response to the overall errors in PS projections. This implies that differences in regional forcings are more important in causing different climate change patterns that standard PS cannot emulate, though nonlinear behavior may be more important for scenarios that stabilize (Caesar et al. 2013). There is some evidence for nonlinearities contributing to PS errors beyond 4°C global warming, however, and we can attribute these to processes affecting the northern Canadian coast (e.g., sea ice feedbacks) and regions of South America (moisture dynamics) where we see that errors, in the worst performing patterns, are more apparent *and* nonlinear response to CO_2 is strongest.

Acknowledgments. The European Commission's 7th Framework Programme (EU/FP7) under Grant 603864 (HELIX) supported all authors. T. J. O. and C. J. W. were also supported by the TOPDAD project

(308620-ENV.2012.6.1-3). We acknowledge the Met Office Hadley Centre and the U.S. Department of Energy's Program for Climate Model Diagnosis and Intercomparison (PCMDI) for producing and making available the HadGEM2-ES climate model data used in this study. We thank two reviewers for suggestions that improved this paper.

REFERENCES

Arnell, N. W., and Coauthors, 2013: A global assessment of the effects of climate policy on the impacts of climate change. *Nat. Climate Change*, **3**, 512–519, https://doi.org/10.1038/nclimate1793.

——, J. A. Lowe, B. Lloyd-Hughes, and T. J. Osborn, 2018: The impacts avoided with a 1.5°C climate target: A global and regional assessment. *Climatic Change*, **147**, 61–76, https://doi.org/10.1007/s10584-017-2115-9.

Caesar, J., E. Palin, S. Liddicoat, J. Lowe, E. Burke, A. Pardaens, M. Sanderson, and R. Kahana, 2013: Response of the HadGEM2 Earth system model to future greenhouse gas emissions pathways to the year 2300. *J. Climate*, **26**, 3275–3284, https://doi.org/10.1175/JCLI-D-12-00577.1.

Collins, W. J., and Coauthors, 2011: Development and evaluation of an Earth-system model—HadGEM2. *Geosci. Model Dev.*, **4**, 1051–1075, https://doi.org/10.5194/gmd-4-1051-2011.

Good, P., and Coauthors, 2015: Nonlinear regional warming with increasing CO_2 concentrations. *Nat. Climate Change*, **5**, 138–142, https://doi.org/10.1038/nclimate2498.

Gosling, S. N., and N. W. Arnell, 2016: A global assessment of the impact of climate change on water scarcity. *Climatic Change*, **134**, 371–385, https://doi.org/10.1007/s10584-013-0853-x.

Heinke, J., S. Ostberg, S. Schaphoff, K. Frieler, C. Müller, D. Gerten, M. Meinshausen, and W. Lucht, 2013: A new climate dataset for systematic assessments of climate change impacts as a function of global warming. *Geosci. Model Dev.*, **6**, 1689–1703, https://doi.org/10.5194/gmd-6-1689-2013.

Ishizaki, N. N., H. Shiogama, K. Takahashi, S. Emori, K. Dairaku, H. Kusaka, T. Nakaegawa, and I. Takayabu, 2012: An attempt to estimate of probabilistic regional climate analogue in a warmer Japan. *J. Meteor. Soc. Japan*, **90B**, 65–74, https://doi.org/10.2151/jmsj.2012-B05.

Jones, C. D., and Coauthors, 2011: The HadGEM2-ES implementation of CMIP5 centennial simulations. *Geosci. Model Dev.*, **4**, 543–570, https://doi.org/10.5194/gmd-4-543-2011.

Lynch, C., C. Hartin, B. Bond-Lamberty, and B. Kravitz, 2017: An open-access CMIP5 pattern library for temperature and precipitation: Description and methodology. *Earth Syst. Sci. Data*, **9**, 281–292, https://doi.org/10.5194/essd-9-281-2017.

Mitchell, J. F. B., T. C. Johns, M. Eagles, W. J. Ingram, and R. A. Davis, 1999: Towards the construction of climate change scenarios. *Climatic Change*, **41**, 547–581, https://doi.org/10.1023/A:1005466909820.

Mitchell, T. D., 2003: Pattern scaling: An examination of the accuracy of the technique for describing future climates. *Climatic Change*, **60**, 217–242, https://doi.org/10.1023/A:1026035305597.

Morice, C. P., J. J. Kennedy, N. A. Rayner, and P. D. Jones, 2012: Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 data set. *J. Geophys. Res.*, **117**, D08101, https://doi.org/10.1029/2011JD017187.

Osborn, T. J., C. J. Wallace, I. C. Harris, and T. M. Melvin, 2016: Pattern scaling using ClimGen: Monthly-resolution future climate scenarios including changes in the variability of precipitation. *Climatic Change*, **134**, 353–369, https://doi.org/10.1007/s10584-015-1509-9.

Ostberg, S., W. Lucht, S. Schaphoff, and D. Gerten, 2013: Critical impacts of global warming on land ecosystems. *Earth Syst. Dyn.*, **4**, 347–357, https://doi.org/10.5194/esd-4-347-2013.

Sgubin, G., D. Swingedouw, S. Drijfhout, S. Hagemann, and E. Robertson, 2015: Multimodel analysis on the response of the AMOC under an increase of radiative forcing and its symmetrical reversal. *Climate Dyn.*, **45**, 1429–1450, https://doi.org/10.1007/s00382-014-2391-2.

Taylor, K. E., R. J. Stouffer, and G. A. Meehl, 2012: An overview of CMIP5 and the experiment design. *Bull. Amer. Meteor. Soc.*, **93**, 485–498, https://doi.org/10.1175/BAMS-D-11-00094.1.

Tebaldi, C., and J. M. Arblaster, 2014: Pattern scaling: Its strengths and limitations, and an update on the latest model simulations. *Climatic Change*, **122**, 459–471, https://doi.org/10.1007/s10584-013-1032-9.

van Vuuren, D. P., B. Eickhout, P. L. Lucas, and M. G. J. den Elzen, 2006: Long-term multi-gas scenarios to stabilise radiative forcing—Exploring costs and benefits within an integrated assessment framework. *Energy J.*, **27**, 10, https://doi.org/10.5547/ISSN0195-6574-EJ-VolSI2006-NoSI3-10.

——, and Coauthors, 2011: The representative concentration pathways: An overview. *Climatic Change*, **109**, 5–31, https://doi.org/10.1007/s10584-011-0148-z.

Warren, R., and Coauthors, 2008: Development and illustrative outputs of the community integrated assessment system (CIAS), a multi-institutional modular integrated assessment approach for modelling climate change. *Environ. Modell. Software*, **23**, 592–610, https://doi.org/10.1016/j.envsoft.2007.09.002.

——, and Coauthors, 2013: The AVOID programme's new simulations of the global benefits of stringent climate change mitigation. *Climatic Change*, **120**, 55–70, https://doi.org/10.1007/s10584-013-0814-4.