Climate Model Dependence and the Ensemble Dependence Transformation of CMIP Projections

G. ABRAMOWITZ

ARC Centre of Excellence for Climate System Science, and Climate Change Research Centre,
University of New South Wales, Sydney, New South Wales, Australia

C. H. BISHOP

Naval Research Laboratory, Monterey, California

(Manuscript received 20 May 2014, in final form 26 September 2014)

ABSTRACT

Obtaining multiple estimates of future climate for a given emissions scenario is key to understanding the likelihood and uncertainty associated with climate-related impacts. This is typically done by collating model estimates from different research institutions internationally with the assumption that they constitute independent samples. Heuristically, however, several factors undermine this assumption: shared treatment of processes between models, shared observed data for evaluation, and even shared model code. Here, a ‘‘perfect model’’ approach is used to test whether a previously proposed ensemble dependence transformation (EDT) can improve twenty-first-century Coupled Model Intercomparison Project (CMIP) projections. In these tests, where twenty-first-century model simulations are used as out-of-sample ‘‘observations,’’ the mean-square difference between the transformed ensemble mean and ‘‘observations’’ is on average 30% less than for the untransformed ensemble mean. In addition, the variance of the transformed ensemble matches the variance of the ensemble mean about the ‘‘observations’’ much better than in the untransformed ensemble. Results show that the EDT has a significant effect on twenty-first-century projections of both surface air temperature and precipitation. It changes projected global average temperature increases by as much as 16% (0.2°C for B1 scenario), regional average temperatures by as much as 2.6°C (RCP8.5 scenario), and regional average annual rainfall by as much as 410 mm (RCP6.0 scenario). In some regions, however, the effect is minimal. It is also found that the EDT causes changes to temperature projections that differ in sign for different emissions scenarios. This may be as much a function of the makeup of the ensembles as the nature of the forcing conditions.

1. Introduction

The Coupled Model Intercomparison Project (CMIP) climate model ensembles (Meehl et al. 2007a; Taylor et al. 2012) that underpin Intergovernmental Panel on Climate Change (IPCC) assessment reports (Meehl et al. 2007b; CMIP 2013a) are commonly described as ‘‘ensembles of opportunity’’ (Tebaldi and Knutti 2007; Knutti et al. 2010a), in the sense that their makeup is determined by the ability of climate research groups to contribute to them. Scientific ability is typically mitigated by operational pressures, funding limitations, or computational resource constraints, so that submissions from any particular group, should they be able to participate, generally reflect their institution’s internal priorities.

As a sampling strategy for projecting possible climate futures, this approach is extremely difficult to disentangle. Some groups may submit a single simulation for a given emissions scenario, others an entire ensemble (generated in a variety of different ways) or even several ensembles from variants of their own model (see CMIP 2013b). In addition, since research groups share expertise, a literature base, observational datasets, and even model code, the probability that climate models share systematic errors is high and hence the assumption of model independence is correspondingly poor (S. Jewson and E. Hawkins 2009, unpublished manuscript; Abramowitz 2010; Knutti et al. 2010a,b).

Corresponding author address: Gab Abramowitz, ARC Centre of Excellence for Climate System Science, and Climate Change Research Centre, University of New South Wales, Sydney NSW 2052, Australia.
E-mail: gabriel@unsw.edu.au

DOI: 10.1175/JCLI-D-14-00364.1

Any quantitative definition of dependence in this context, however, requires assumptions about the relationship between observations and model ensemble spread, as well as what causes the spread. If, for example, we believe that a ''perfect model'' should essentially match observations of a particular variable, excepting some noise from model approximations—the so-called truth-plus-error or truth-centered paradigm of interpretation (Knutti et al. 2010a; Annan and Hargreaves 2010)—then truly independent models should have zero error correlation, as we would expect for independent random variables.

While this conceptualization of an ensemble of independent models is intuitive, it is not appropriate for ensembles of the climate system. If errors (i.e., model minus observation) from independent models were uncorrelated, the error variance of the ensemble mean would be inversely proportional to the ensemble size, going to zero as the ensemble size approaches infinity. The only barrier to the ensemble mean approximating observations arbitrarily closely in this case is the number of independent models in the ensemble. Aside from there being strong anecdotal evidence contradicting this paradigm of ensemble interpretation (e.g., Knutti et al. 2010b; Annan and Hargreaves 2010), it implies that internal variability, such as El Niño–Southern Oscillation (ENSO) and even weather patterns, is predictable with arbitrary precision. That is, the truth-plus-error paradigm of interpretation implies that the climate system is deterministic.

While it might be tempting to argue that each model's internal variability should be considered as part of the ''model noise'' within the truth-plus-error paradigm, the internal variability in the observations will be common to the model-minus-observation time series of all models. That is, the model mean will not converge to the observations, and the model-minus-observation time series for independent models will have nonzero correlation. Alternatively, suggesting that the truth-plus-error paradigm applies only to long time scale averages is to ignore the very high likelihood that internal variability operates on longer time scales (e.g., Ault et al. 2013; James and James 1989). In short, we see no defensible justification for the truth-plus-error paradigm.

An alternative is to conceptualize models and observations as being drawn from the same distribution (e.g., Annan and Hargreaves 2010; Bishop and Abramowitz 2013). Imagine, for example, a distribution defined by sampling across many replicates of Earth under identical climate forcing and different but observationally consistent initial conditions. Chaotic aspects of the climate system would then naturally lead to a range of possible climate system states in the replicate Earths. This idea is already well accepted in the context of weather prediction (Hamill et al. 2000; Gneiting and Raftery 2005) and climate models' internal variability (Collins et al. 2001; Deser et al. 2012). In this conceptualization of a perfect ensemble, our own Earth would effectively be one random sample from the climate probability distribution function (CPDF) defined by these replicate Earths and a perfect model simulation, were we able create such a thing, would be another random sample (Bishop and Abramowitz 2013). The CPDF would therefore define the nature of the internal variability of Earth's climate system and allow true probabilistic prediction of greenhouse gas–induced change for a given emissions scenario, not just in mean temperatures but also the frequencies of high impact events such as droughts, floods, heatwaves, storm surges, and tropical cyclones.

We clearly cannot claim that climate models are replicate Earths in this sense, but viewing them as attempts to create replicate Earths is useful. While we may conceptually understand perturbed initial conditions ensembles as being attempts to define the CPDF, there is evidence that they exhibit too little internal variability (Haughton et al. 2014; Ault et al. 2013; England et al. 2014), and at least heuristically most would agree that an ensemble generated by a single model is insufficient for producing independent estimates of future climate. Dependence within an ensemble increases the discrepancy between the PDF of climate model forecasts and the true CPDF associated with a prescribed emissions scenario.

In attempting to ameliorate the ill effects of model dependence on ensemble-based estimates of the CPDF, Bishop and Abramowitz (2013) showed that an ensemble of replicate Earths (perfect models) would have two key statistical properties:

1) The best estimate to any replicate Earth (in a mean-square difference sense) is the mean of the CPDF. Over a long enough time period, the linear combination of replicate Earths that would minimize the mean-square distance to observations of any variable on our Earth would be the equally weighted ensemble mean.
2) Over a long enough period, all replicate Earths would have the same variance about the CPDF mean. In particular, if the temporal change in the CPDF variance is slow, the time-averaged mean of the instantaneous CPDF variance will approximate the variance over time of our Earth about the CPDF mean.

Bishop and Abramowitz (2013) developed a mathematical postprocessing procedure that transforms raw dependent ensembles into ensembles that have the above two replicate Earth-like properties. We will refer to this as the ensemble dependence transformation (EDT). They also showed that the EDT resulted in much flatter rank histograms for surface air temperature when compared to 1970–99 Hadley Centre/Climate Research Unit Temperature data, version 3 (HadCRUT3; Brohan et al. 2006), suggesting that observations and the transformed ensemble are more likely to be drawn from the same distribution than observations and the original CMIP ensemble. The work presented here aims to test the efficacy of applying this approach to improve twenty-first-century CMIP projections and examines the results of doing so.

In the next section, we briefly describe the model simulations and observational data used to investigate the effect of the EDT. In section 3, we review the workings of the EDT process. In section 4, we create an out-of-sample testing environment to try to examine the ability of the EDT to improve twenty-first-century CMIP projections. In section 5, we examine the results of applying the EDT to CMIP5 surface air temperature and precipitation projections. Section 6 contains discussion and conclusions.

2. Model simulations and observational data

The EDT process uses properties 1 and 2 above, together with an observational dataset and a twentieth-century CMIP ensemble, to try to infer characteristics of the true CPDF. We examine surface air temperature using the monthly 5° × 5° HadCRUT4.2 dataset (Morice et al. 2012) for the period 1970–2004 and precipitation using the monthly Global Precipitation Climatology Project, version 2.2 (GPCP2.2) 2.5° × 2.5° dataset (Adler et al. 2003) for 1979–2004. Parameters for the EDT are derived for both phase 5 of CMIP (CMIP5; Taylor et al. 2012) and phase 3 of CMIP (CMIP3; Meehl et al. 2007a) ensembles using all historical simulations that have contiguous twenty-first-century counterparts for each scenario. We include representative concentration pathways (RCP2.6, RCP4.5, RCP6.0, and RCP8.5; Meinshausen et al. 2011) and SRES scenarios (B1, A2, and A1B; Nakicenovic and Swart 2000). Note that for the three CMIP3 ensembles the in-sample period is shortened to 1970–99 (temperature) and 1979–99 (precipitation). The CMIP5 projection period was shortened to 2006–99 and CMIP3 was shortened to 2004–98 to maximize the number of simulations available. All parameter calculations weight grid cells by their relative surface areas. The number of contiguous (e.g., 1970–2099) simulations for each model and each scenario is shown in Table 1. Once derived, the EDT parameters are applied out of sample to each twenty-first-century scenario ensemble. The process, described in detail below, is summarized in the flowchart in Fig. 1.

TABLE 1. The number of simulations from each CMIP5 model used that had contiguous historical and RCP simulations, shown separately for each RCP. Rows with boldface numbers indicate models that contributed different numbers of runs to each RCP, affecting the comparability of ensemble results for different RCPs. Italicized rows indicate models that were excluded from the pseudo-independent sampling strategy. Model acronym expansions can be found at CMIP (2013b) and also at http://www.ametsoc.org/PubsAcronymList.

Model name	RCP2.6	RCP4.5	RCP6.0	RCP8.5
ACCESS1.0	**0**	**1**	**0**	**1**
ACCESS1.3	***0***	***1***	***0***	***1***
BCC_CSM1.1	1	1	1	1
BCC_CSM1.1(m)	*1*	*1*	*1*	*1*
BNU-ESM	**1**	**1**	**0**	**1**
CanESM2	**4**	**4**	**0**	**4**
CCSM4	6	6	6	6
CESM1(BGC)	***0***	***1***	***0***	***1***
CESM1(CAM5)	**3**	**3**	**0**	**3**
CESM1(WACCM)	***0***	***1***	***0***	***1***
CMCC-CESM	***0***	***0***	***0***	***1***
CMCC-CM	***0***	***1***	***0***	***0***
CMCC-CMS	**0**	**1**	**0**	**1**
CNRM-CM5	**1**	**1**	**0**	**5**
CSIRO Mk3.6.0	10	10	10	10
EC-EARTH	**2**	**7**	**0**	**6**
FGOALS-g2	***1***	***1***	***0***	***1***
FGOALS-s2	**1**	**3**	**1**	**3**
FIO-ESM	3	3	3	3
GFDL CM3	1	1	1	1
GFDL-ESM2G	*1*	*1*	*1*	*1*
GFDL-ESM2M	*1*	*1*	*1*	*1*
GISS-E2-H	**3**	**13**	**3**	**3**
GISS-E2-H-CC	***0***	***1***	***0***	***0***
GISS-E2-R	***3***	***17***	***1***	***3***
GISS-E2-R-CC	***0***	***1***	***0***	***0***
HadGEM2-AO	*1*	*1*	*1*	*1*
HadGEM2-CC	***0***	***1***	***0***	***3***
HadGEM2-ES	**4**	**4**	**3**	**4**
INM-CM4	**0**	**1**	**0**	**1**
IPSL-CM5A-LR	***4***	***4***	***1***	***4***
IPSL-CM5A-MR	1	1	1	1
IPSL-CM5B-LR	***0***	***1***	***0***	***1***
MIROC-ESM	*1*	*1*	*1*	*1*
MIROC-ESM-CHEM	*1*	*1*	*1*	*1*
MIROC5	**3**	**3**	**1**	**3**
MPI-ESM-LR	***3***	***3***	***0***	***3***
MPI-ESM-MR	**1**	**3**	**0**	**1**
MRI-CGCM3	1	1	1	1
NorESM1-M	1	1	1	1
NorESM1-ME	*1*	*1*	*1*	*1*
Pseudo-independent	**18**	**21**	**12**	**21**
Total simulations	**65**	**109**	**41**	**86**

3. The ensemble dependence transformation

Since we only have one Earth, categorically defining the properties of the true CPDF is impossible. By

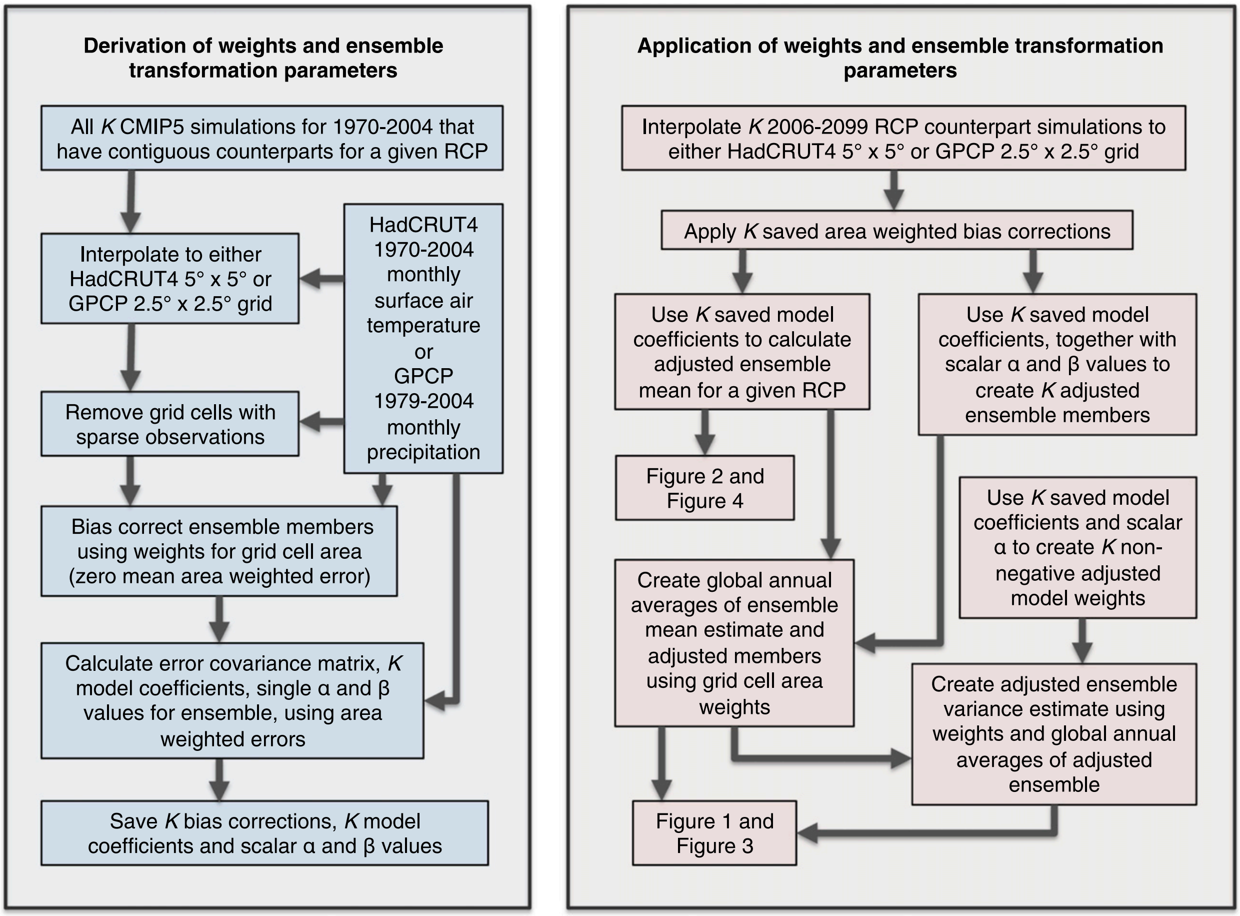
Derivation of weights and ensemble transformation parameters
All K CMIP5 simulations for 1970-2004 that have contiguous counterparts for a given RCP
Interpolate to either HadCRUT4 5° x 5° or GPCP 2.5° x 2.5° grid
HadCRUT4 1970-2004 monthly surface air temperature or GPCP 1979-2004 monthly precipitation
Remove grid cells with sparse observations
Bias correct ensemble members using weights for grid cell area (zero mean area weighted error)
Calculate error covariance matrix, K model coefficients, single α and β values for ensemble, using area weighted errors
Save K bias corrections, K model coefficients and scalar α and β values
Application of weights and ensemble transformation parameters
Interpolate K 2006-2099 RCP counterpart simulations to either HadCRUT4 5° x 5° or GPCP 2.5° x 2.5° grid
Apply K saved area weighted bias corrections
Use K saved model coefficients to calculate adjusted ensemble mean for a given RCP
Use K saved model coefficients, together with scalar α and β values to create K adjusted ensemble members
Figure 2 and Figure 4
Use K saved model coefficients and scalar α to create K non-negative adjusted model weights
Create global annual averages of ensemble mean estimate and adjusted members using grid cell area weights
Create adjusted ensemble variance estimate using weights and global annual averages of adjusted ensemble
Figure 1 and Figure 3

FIG. 1. Step-by-step procedure for deriving and applying the ensemble dependence transformation (EDT).

comparing climate model ensembles with observations, we can, however, better utilize the information that they provide toward estimating CPDF properties than simply assuming that raw ensemble members are samples from the true CPDF.

The first step in this process is to obtain an estimate of the (time varying) CPDF mean. To do this, we will first show that we can derive a linear combination of the twentieth-century CMIP simulations that is closer to observations (in a mean-square difference sense) than their equally weighted mean. This shows that (i) the CMIP ensemble in not replicate Earth like (it does not satisfy 1 above) and (ii) the weighted mean we derive is a better estimate of the CPDF mean than the equally weighted mean. In statistically estimating this linear combination, one avoids "overfitting" the in-sample data by ensuring that one has many more observations in the training dataset than the number of models used to fit the data.

All of the calculations below assume that each simulation has been bias corrected, in the sense that each simulation and observations have the same mean over the entire space and time domain in the in-sample period (i.e., 1970–2004 or 1979–2004). While there are compelling reasons not to do this (Ehret et al. 2012), it remains standard practice in the climate community (e.g., Meehl et al. 2007b) and greatly simplifies the solutions to the problems posed below. Key results below follow Bishop and Abramowitz (2013), where they are covered in detail; here, we only briefly summarize the main points for completeness.

Suppose we wish to find the linear combination of model simulations that minimizes mean-square difference with respect to observations. That is,

$$\mu_e^j = \sum_{k=1}^{K} w_k x_k^j \quad \text{so that} \quad \sum_{j=1}^{J} (\mu_e^j - y^j)^2$$

is minimized, where x_k^j is the jth space–time step of the kth model simulation and y^j is the jth space–time step of the observational dataset (where J is the number of grid cells multiplied by the number of months in the

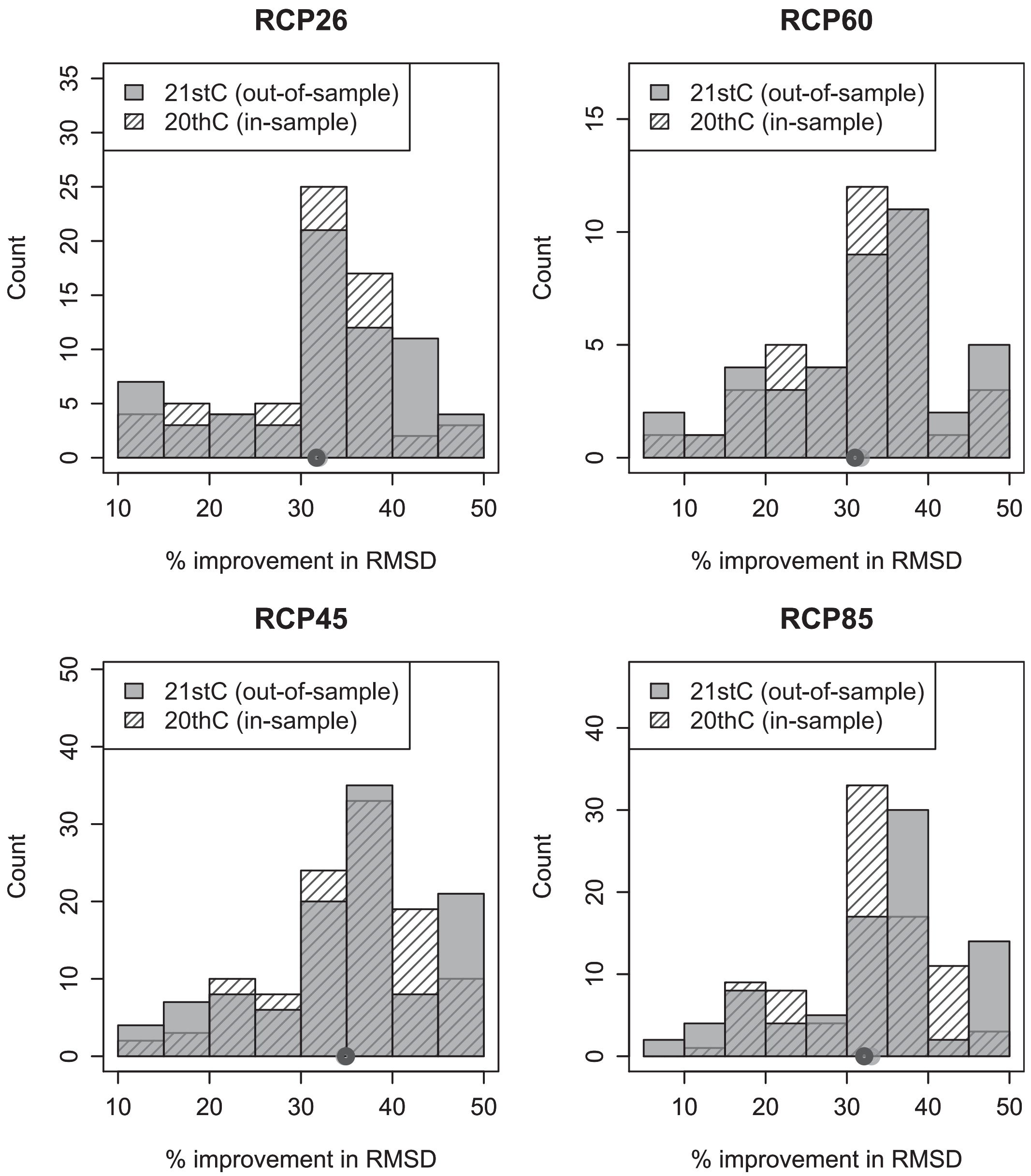

FIG. 2. Histograms of percentage improvement in RMSD in ensemble mean surface air temperature (over time and space; weighted for gridcell area) afforded by the EDT when tested out of sample in the twenty-first century, for each RCP. Histograms are collated over all possible simulations as "observations" for each RCP. Means are shown on the horizontal axis.

in-sample period). If we additionally require the coefficients w_k sum to 1, we can obtain an analytical solution using a Lagrange multiplier. Each w_k in this solution is proportional to the sum of the kth row of the inverse of the symmetric matrix

$$\mathbf{A} = \begin{bmatrix} \sigma_{1,1}^2 & \cdots & \sigma_{1,K}^2 \\ \vdots & \ddots & \vdots \\ \sigma_{K,1}^2 & \cdots & \sigma_{K,K}^2 \end{bmatrix},$$

where $\sigma_{i,j}^2$ is the "error" covariance between the ith and jth simulations. This result is reported in both Potempski and Galmarini (2009) and Bishop and Abramowitz (2013). Note that the solution for the kth coefficient accounts for the performance of the kth simulation (since $\sigma_{k,k}^2$ is the kth simulation "error" variance) and depends heavily on the "error" covariance between the kth and other simulations: precisely what one might heuristically expect for weights that account for dependence (e.g., Jun et al. 2008).

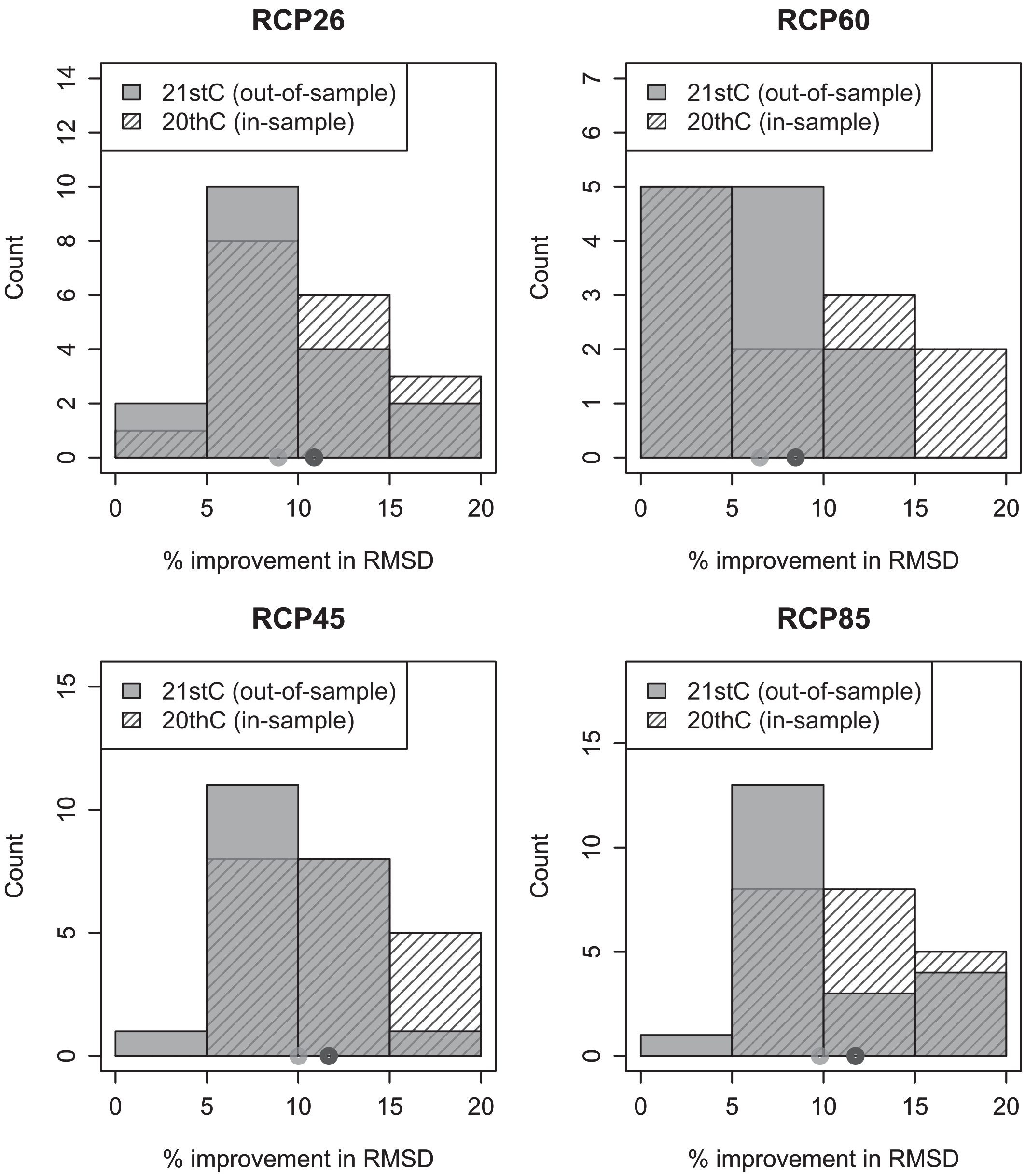

FIG. 3. As in Fig. 2, but only using members of the pseudo-independent ensemble.

Note that "error" is in quotations here since a significant component of model–observation difference is a result of chaotic internal system variability, rather than error per se.

Since the coefficients w_k are an analytical solution to the problem posed above, the linear combination μ_e is by definition the best possible linear combination of the simulations at hand, at least for the in-sample period (we will test it out of sample in the next section). We therefore choose μ_e as our CPDF mean estimate.

The next step is to estimate the variance of the CPDF. Using our CPDF mean estimate obtained above, we calculate the sample variance of observations about this CPDF mean estimate for the in-sample period,

$$s_e^2 = \frac{\sum_{j=1}^{J} (\mu_e^j - y^j)^2}{J - 1}.$$

While this variance is calculated over time, rather than across an ensemble, it does provide us with an estimate of the time average of instantaneous CPDF variance over the in-sample period (following property 2 above), if the rate of change of true instantaneous CPDF variance is

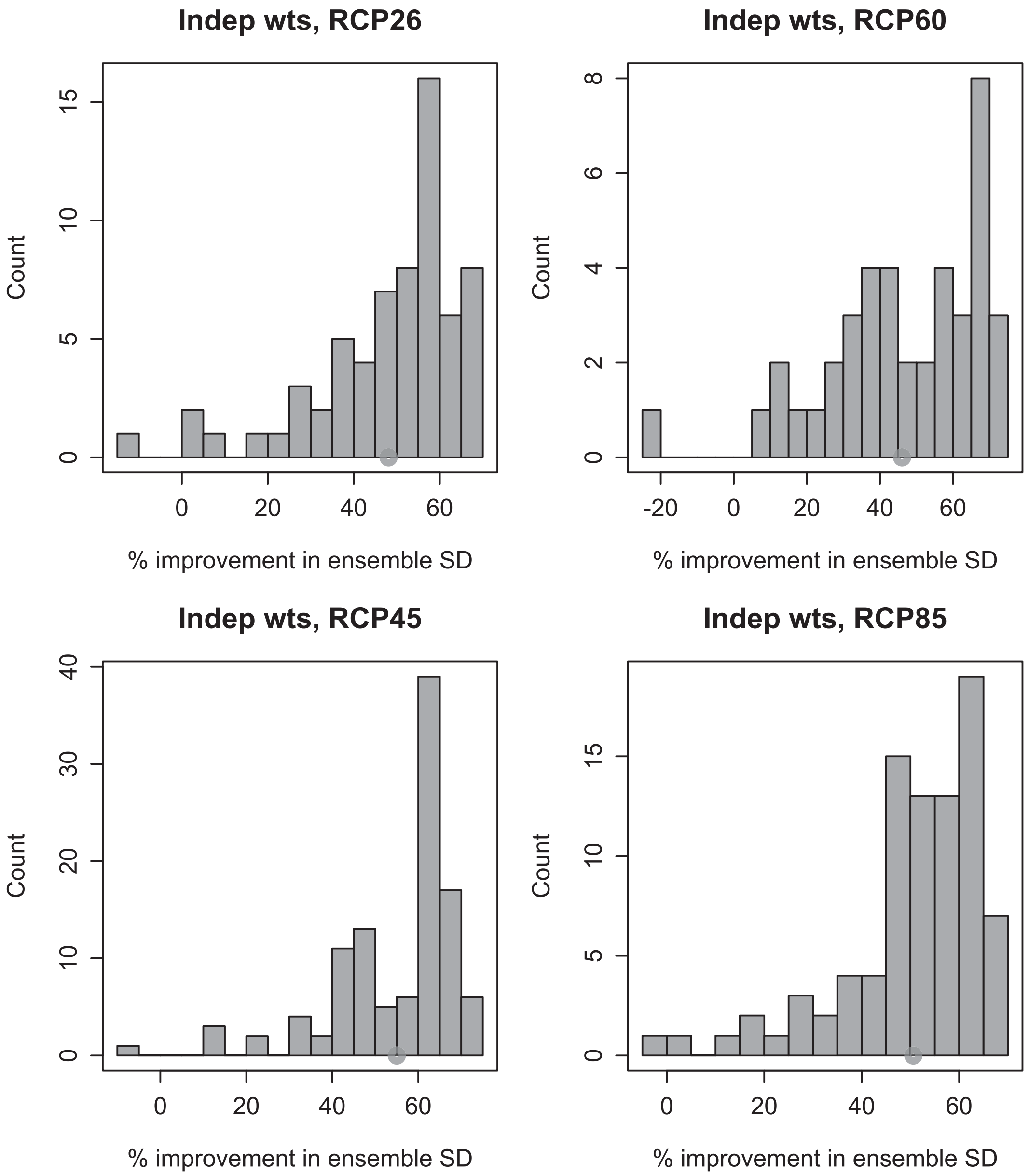

FIG. 4. Histograms of percentage improvement in ensemble standard deviation for surface air temperature (over time and space; weighted for gridcell area) afforded by the EDT when tested out of sample in the twenty-first century, for each RCP. Histograms are collated over all possible simulations as ‘‘observations’’ for each RCP. Means are shown on the horizontal axis.

slow. If, for example, instantaneous CPDF variance were static over the in-sample period, s_e^2 would provide a very accurate estimate of instantaneous CPDF variance.

If we had a true replicate Earth ensemble, we would expect s_e^2 to be approximately equal to the average of CPDF variance at each grid cell at each time step in the in-sample period,

$$s_e^2 \cong \frac{1}{J}\sum_{j=1}^{J}\sigma_e^{2j}.$$

The fact that the CMIP ensembles do not satisfy this equation shows that they do not satisfy property 2 above. To transform them so that they do, we define

$$\tilde{w}_k = \frac{[w_k + (\alpha - 1)/K]}{\alpha} \quad \text{and}$$

$$\tilde{x}_k^j = \mu_e^j + \beta[(\bar{x}^j + \alpha x'^j_k) - \mu_e^j],$$

where $\alpha = 1 - K\min(w_k)$ and $\min(w_k)$ is the lowest (most negative) w_k (note that the minimization problem above did not ensure positive w_k). In the second expression, $\bar{x}^j$ is the multimodel mean for the jth space–time step, $x_k'^j$ is the kth simulation's deviation from that mean, and β is a scalar parameter that ensures that s_e^2 is equal to the average of CPDF variance, as described above [see Bishop and Abramowitz (2013) for more detail]. These linear transformations ensure that (i) the $\tilde{w}$ still sum to 1 and are now all positive, (ii) the minimum "error" variance estimate is preserved,

$$\mu_e^j = \sum_{k=1}^{K} w_k x_k^j = \sum_{k=1}^{K} \tilde{w}_k \tilde{x}_k^j,$$

and (iii) the variance condition described above,

$$s_e^2 \cong \frac{1}{J} \sum_{j=1}^{J} \sum_{k=1}^{K} \tilde{w}_k (\tilde{x}_k^j - \mu_e^j)^2,$$

is satisfied (where the inner sum is simply a weighted estimate of the instantaneous CPDF variance). Details, including proofs, can be found in Bishop and Abramowitz (2013).

The transformation for the entire ensemble therefore relies only on the original w_k estimated from the minimization of errors problem above and the two scalar constants, α and β. From the expression for $\tilde{x}_k^j$ above, we can see that α acts to expand ensemble variance about the multimodel mean, and β then acts to contract it about the CPDF mean estimate.

With this transformed ensemble we now have a more credible estimate of the CPDF mean μ_e^j and of CPDF variance $\sigma_e^{2j} = \sum_{k=1}^{K} \tilde{w}_k (\tilde{x}_k^j - \mu_e^j)^2$. To apply these to CMIP projections, the w_k, α, and β are derived on the in-sample twentieth-century data for each scenario subset and then applied to the appropriate projection (see Fig. 1). The derivation of these weights, the transformation parameters, and global means described below all use area weighting to account for the range of surface areas represented by each grid cell in the 5° × 5° HadCRUT4 and 2.5° × 2.5° GPCP2.2 grids.

4. Can we trust the EDT on twenty-first-century projections?

Bishop and Abramowitz (2013) showed that the EDT described above resulted in much flatter rank histograms for surface air temperature, suggesting that observations and the transformed ensemble are more likely to be drawn from the same distribution than observations and the original CMIP ensemble. While this reassures us that ensemble variance estimates are improved, can we have any confidence that applying the process to the twenty-first-century CMIP projections will yield improved results? Is the twenty-first century likely to be different enough from the twentieth century that the EDT may in fact degrade projections rather than improve them? Are the dependence weights $\tilde{w}_k$ favoring models whose internal variability happens to coincide with observational variability (e.g., El Niño–Southern Oscillation or Indian Ocean dipole phenomena) or simply fitting to noise?

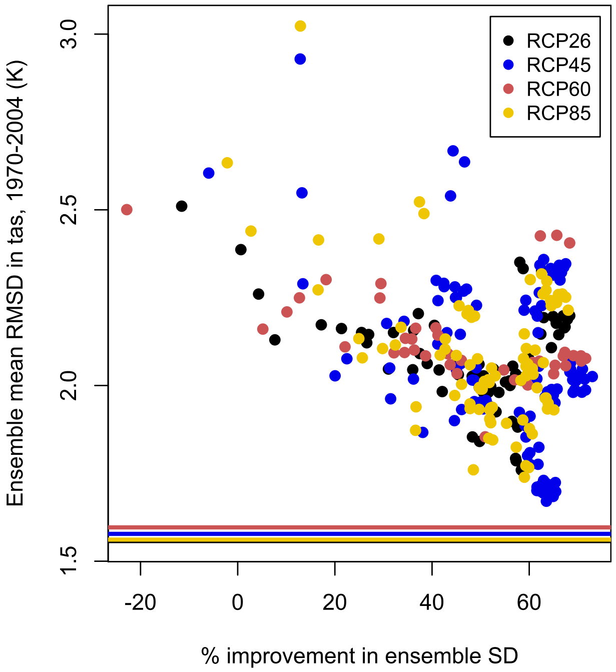

FIG. 5. Relationship between percentage improvement in ensemble standard deviation, as defined for Fig. 4, and ensemble RMSD over space and time in the in-sample period (1970–2004). Results are shown for surface air temperature (tas). RMSD values for each in-sample RCP subset ensemble compared to HadCRUT4 are shown by horizontal lines.

To address these questions, we construct a model-as-truth or perfect model experiment. This involves nominating a single model simulation (i.e., a contiguous twentieth-century + RCP simulation for 1970–2099) to be treated as "observations" and using the remaining simulations as the raw ensemble. The parameters defining the EDT of this raw ensemble are defined by "observations" from the in-sample 1970–2004 period. The performance of the transformed ensemble is then tested out of sample by comparing it to "observations" from 2006 to 2099. This approach is designed to test the ability of the EDT approach by effectively providing "observations" of the twenty-first century. It allows us to directly test (i) whether the transformed ensemble's mean lies closer to the "observations" than the raw multimodel mean in the out-of-sample period 2006–99 and (ii) whether the transformed ensemble variance is closer to the estimated CPDF variance in the out-of-sample period than the raw

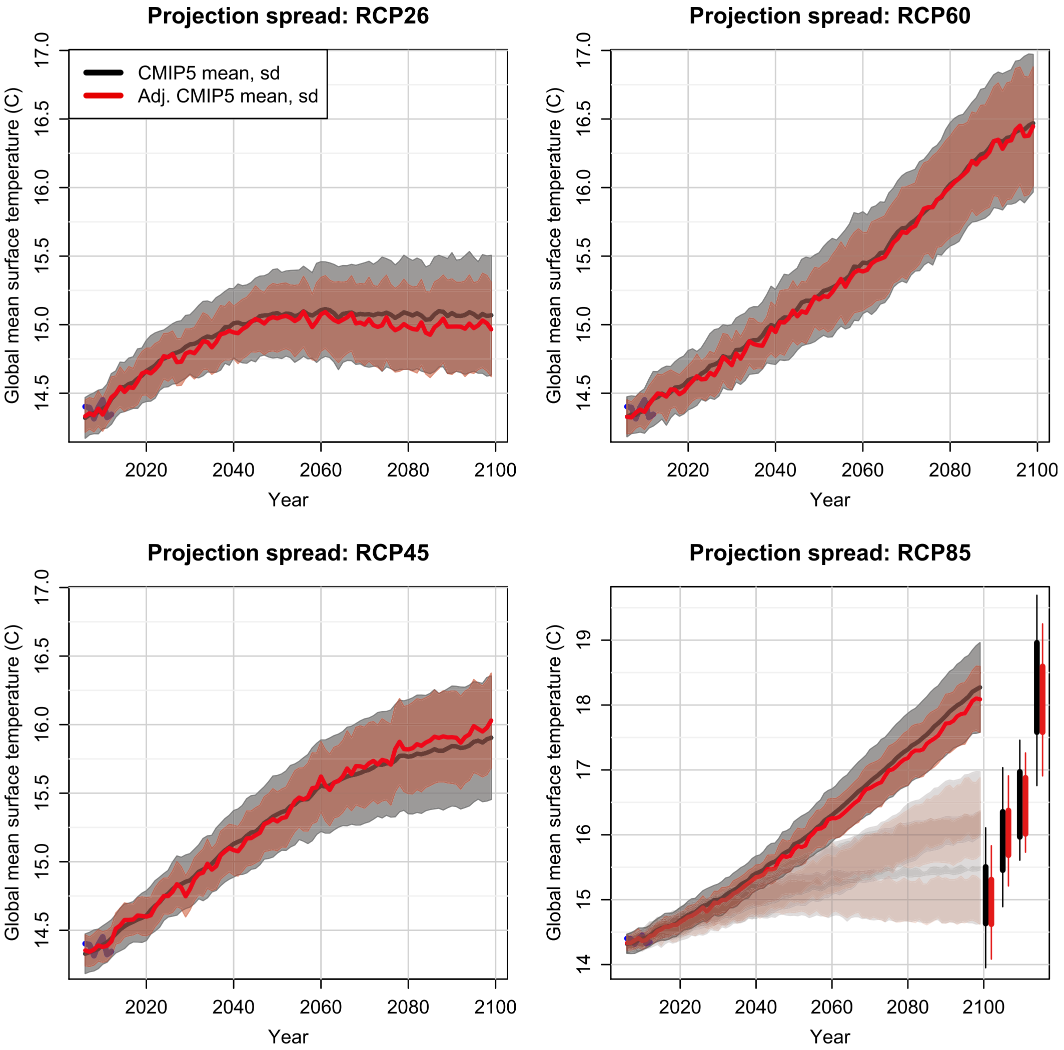

FIG. 6. Projected global surface temperature mean and standard deviation of the CMIP5 ensemble for the original ensemble (black and gray) and EDT ensemble (red). Results are shown for all four RCPs, with observed values for 2005–12 shown in blue.

ensemble variance. The process is repeated for every simulation as "observations" to get a reasonable statistical representation of how this strategy performs out of sample in an environment as similar as possible to the one we wish to apply it in. Collated results are shown in Figs. 2–4.

Figure 2 shows histograms of the percentage improvement in root-mean-square difference (RMSD; over time and space) between the dependence-weighted and unweighted ensemble mean. Results are shown for both in-sample (1970–2004; shown in hatched bars) and out-of-sample (2006–99; gray bars) periods. Histograms are collated over every possible simulation as truth (see the bottom row of Table 1 for the number of simulations for each RCP). The in-sample mean is shown in black and the out-of-sample mean is shown in gray on the horizontal axis. Note that we describe root-mean-square difference rather than error since a considerable proportion of model–observation difference is a result of internal variability, rather than model error.

Critically, the in-sample and out-of-sample distributions are indistinguishable. None of these out-of-sample experiments shows the weights degrading performance relative to the unweighted multimodel mean (all values are positive), and the performance gain in the twentieth-century period afforded by the weighted mean is representative of the performance improvements in the twenty-first century. The approach reduces the root-mean-square difference of the multimodel mean in all RCPs by an average of more than 30%. While not shown here, the equivalent experiment with optimal performance weighting (weights inversely proportional to error variance are optimal for RMSE), as opposed to weights that account for “error” covariance within the ensemble, offers improvements that average only 5%–12%. This result confirms that the observational period we use is almost certainly long enough to construct weights that are useful for twenty-first-century prediction.

Part of the success of the EDT in this experiment is due to its ability to account for dependence in the remaining ensemble once the “observation” simulation has been removed, but some success also stems from the similarity of the “observation” simulation and the simulations remaining in the ensemble. This is likely to be dominant in cases when simulations from the same model form the “observation” simulation and members of the remaining ensemble. This is clearly not always the case, however: many models contributed just one simulation to an RCP experiment (see Table 1). If we analyze the subset of results that only include cases where the “observations” were generated by a model with just one simulation in the CMIP5 ensemble for a given RCP, RMSD improvements do indeed drop to an average of 25% in sample and 23% out of sample.

Since different models from the same institution are likely to be more similar to each other than models from differing institutions, we can further restrict our analysis to those results where the “observations” were a simulation from an institution that contributed just a single simulation for a given RCP. We now only have five contributors—BNU-ESM, CNRM-CM5, FGOALS-s2, INM-CM4, and MRI-CGM3—with in-sample performance ranging from 9% to 25% (16% average) and out-of-sample performance ranging from 10% to 19% (13% average). While we now begin to see a distinction between in-sample and out-of-sample performance, it is still clearly true that in every case the EDT provides out of sample improvements in the multimodel mean.

To investigate this further we now repeat the entire experiment with a heuristically pseudo-independent ensemble subset, by only allowing one simulation from each institution. Models excluded from the pseudo-independent ensemble are italicized in Table 1 and included models contribute only one simulation, with the total number of ensemble members shown in the second to last row of Table 1. Figure 3 shows the application of the same model-as-truth experiment to the pseudo-independent ensemble.

While we might anticipate that this pseudo-independent sampling approach would remove any dependence from the remaining ensemble, all RCPs show that the EDT still provides an improvement of around 10% in RMSD. While there is again a distinction between in-sample and out-of-sample performance, improvements out of sample are still very similar to those in sample.

The results above show that weighting for dependence in this way is not simply favoring those models that have

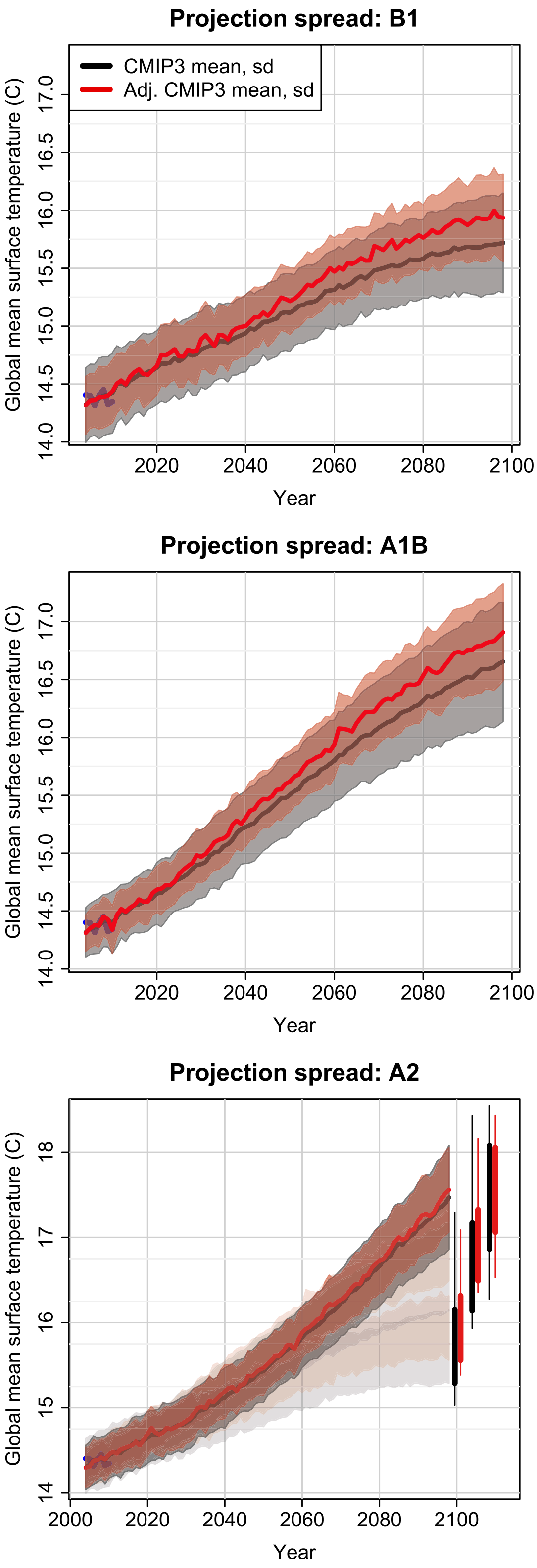

Fig. 7. As in Fig. 6, but for CMIP3.

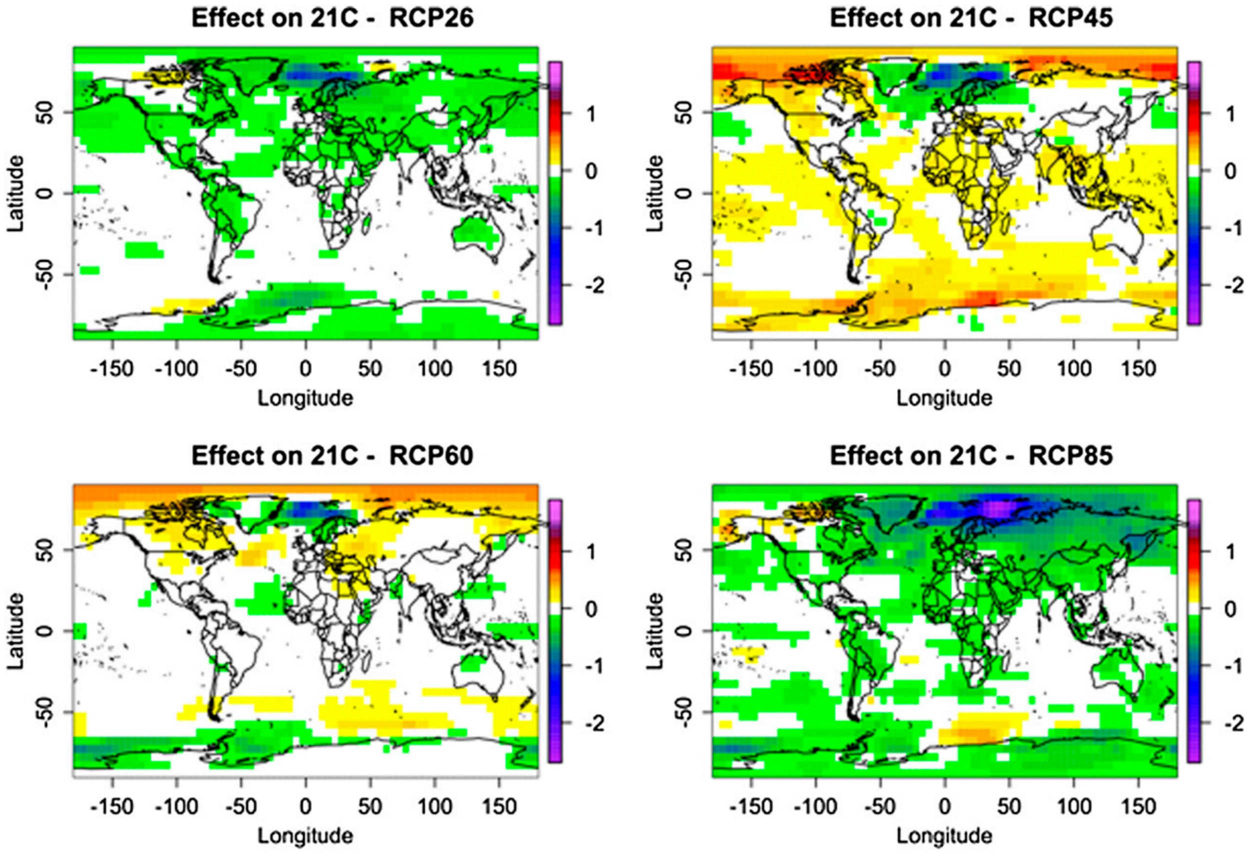

FIG. 8. Regional effect on surface temperature of accounting for dependence within the CMIP5 ensemble, shown for all four RCPs. Values are the differences in the 1980–99 vs 2080–99 surface air temperature change (°C) in the original CMIP5 ensemble mean and the dependence-weighted CMIP5 ensemble mean.

internal variability coincident with observations; rather, if this is the case, the coincidence persists throughout the twenty-first century. The consistent improvements also show the weights are not fitting noise.

To assess the performance of ensemble standard deviation, we first need to establish the target or true standard deviation we wish the EDT process to achieve. In the case of the ensemble mean, this is of course obvious, since we have ‘‘observations’’ for the twenty-first-century out-of-sample period. Recall that we are subscribing to the replicate Earth paradigm, so that the EDT in the in-sample twentieth-century forces ensemble variance, averaged over the in-sample period, to equal the variance of the ‘‘observations’’ about the CPDF mean estimate. It is therefore a natural choice to calculate an equivalent variance for the twenty-first-century period as the target for testing the twentieth-century-derived EDT parameters out of sample. That is, the true CPDF variance estimate for the twenty-first century is calculated by (i) using twenty-first-century models and ‘‘observations’’ to calculate an in-sample twenty-first-century CPDF mean estimate and then (ii) calculating the variance of twenty-first-century ‘‘observations’’ about this CPDF mean estimate. We can then assess whether the EDT parameters derived on twentieth century, applied to the twenty-first-century projections, offer an improvement in ensemble variance over simply using the raw (bias corrected) ensemble mean. This is precisely what is shown in Fig. 4. Histograms are again collated over each possible simulation as truth.

On average, the EDT improves ensemble standard deviation by between 40% and 60%, depending on RCP. This gives us confidence that the transformed ensemble is likely to give a better estimate of the magnitude of internal climate system uncertainty in monthly average temperature in the twenty-first century than using the raw ensemble spread. Better knowledge of the spread of likely system states for a given scenario is clearly valuable for mitigation and adaptation strategies for a range of socioeconomic impacts.

Unlike in the dependence-weighted mean results shown in Fig. 2, each RCP shows one simulation as truth where the EDT did not improve performance in the ensemble variance. Figure 5 shows that in each of these cases, the simulation used as truth had particularly poor performance in RMSE terms and so likely exhibited quite unusual behavior. It shows the percentage improvement in ensemble standard deviation (as in Fig. 4) for a given model as observation on the horizontal axis and the performance of the multimodel mean of the remaining ensemble (measured using RMSE against the model as ‘‘observations’’) on the vertical axis. We use this performance measure as a way of gauging how similar the truth model simulation is to the remaining ensemble. Perhaps unsurprisingly, there is a strong relationship between the similarity of the ‘‘observation’’ simulation and the remaining ensemble and the ability of the EDT to offer ensemble variance improvements. What is more surprising is that the RMSE of each complete in-sample RCP subset ensemble mean when compared against HadCRUT4.2 (shown as horizontal lines) is lower than for any of the model-as-truth cases. That is, the real-world observations and the CMIP5 ensemble mean are more similar in this sense than the CMIP5 ensemble mean (less one simulation) and any one of the simulations that constitute it. These relatively low RMSE values for the real-world data should increase our confidence that the EDT will deliver improvements in ensemble spread out of sample, if the approximately linear nature of the scatter in Fig. 5 gives us any indication.

We also note that Bishop and Abramowitz (2013) explore the out-of-sample performance of the EDT using HadCRUT3 data, including training the EDT on a single decade and testing on others. Again, out-of-sample and in-sample performances were comparable.

5. Results

We now examine the results of applying the EDT shown in Fig. 1 to the CMIP5 and CMIP3 ensembles. Figures 6 and 7 show its effect on globally averaged annual surface air temperature for each of the four CMIP5 RCP and three SRES scenario projections. The original raw ensemble mean is shown in black and the gray shaded region represents one standard deviation across the ensemble. The transformed ensemble mean and standard deviation is shown in red. Note that temperature change in this case is measured relative to a common baseline: the 2005–12 global mean temperature is forced to be the same as observed (blue in Fig. 6) in both cases, as is common practice with model-based change assessments (e.g., Meehl et al. 2007b; Macadam et al. 2010).

The sign and magnitude of the change affected by the EDT is different for different scenario projections. For CMIP5, it induces a −0.10°C global change (−14% warming: i.e., cooler) for RCP2.6 and 0.13°C (8%), −0.02°C (1%), and −0.18°C (5%) change in RCP4.5, RCP6.0, and RCP8.5, respectively. For CMIP3, the EDT results in consistently warmer projections of 0.22° (16%), 0.25° (11%), and 0.09°C (3%) for the B1, A1B, and A2 scenarios, respectively. In both cases, the change in mean appears relatively consistent throughout the twenty-first century, and the variability of the mean has increased to some degree.

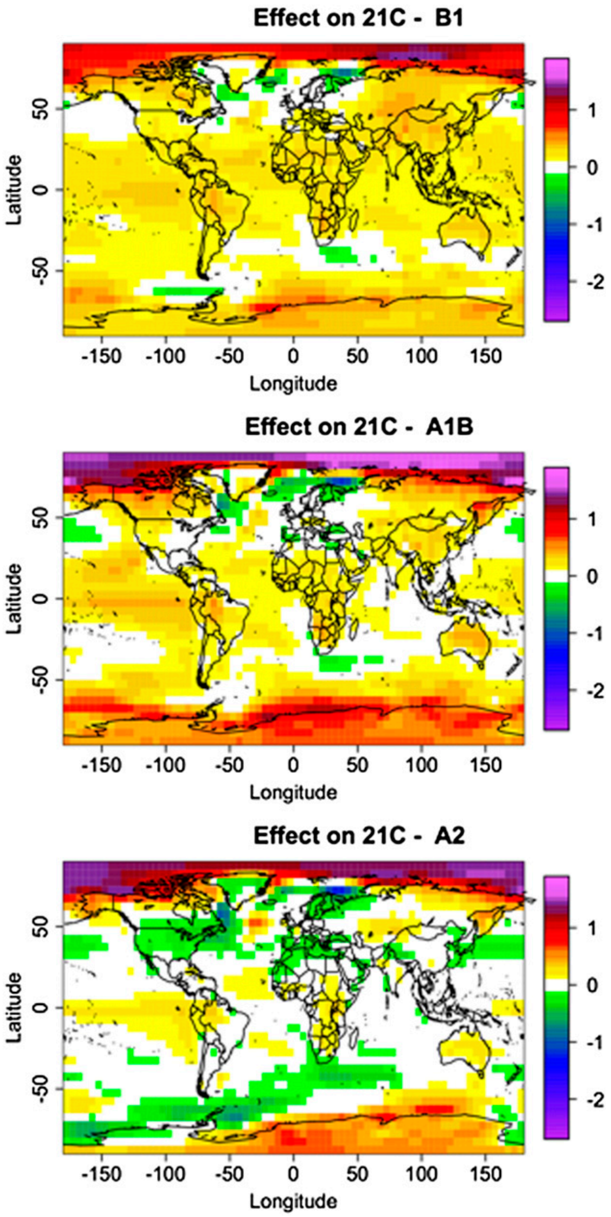

FIG. 9. As in Fig. 8, including color scale, but for CMIP3.

Perhaps more striking in these figures is that the EDT narrows the standard deviation of the ensembles—typically interpreted as uncertainty—in all scenarios, by 12%–26% at the end of the century. This is essentially a result of the EDT adhering to property 2 above. It has already

TABLE 2. The change in various temperature-related quantities as a result of accounting for model simulation dependence within the CMIP5 and CMIP3 ensembles.

	RCP2.6	RCP4.5	RCP6.0	RCP8.5	B1	A1B	A2
Mean global temp 2005–99 (K)	−0.046	+0.017	−0.024	−0.083	+0.117	+0.124	+0.037
Global temp at 2099 (K) (% of warming)	−0.103 (−14%)	+0.125 (+8%)	−0.023 (−1%)	−0.183 (−5%)	+0.217 (+16%)	+0.254 (+11%)	+0.088 (+3%)
Global temp std dev (K) (% increase)	−0.09 (−21%)	−0.105 (−23%)	−0.069 (−14%)	−0.182 (−26%)	−0.052 (−12%)	−0.097 (−19%)	−0.113 (−19%)
Mean Arctic temp diff (K) (70°–90°N)	−0.21	0.21	+0.11	−0.61	+0.54	+0.81	+0.72
Mean Antarctic temp diff (K) (70°–90°S)	−0.11	+0.12	−0.13	−0.26	+0.36	+0.61	+0.22
Peak polar 5° × 5° positive diff (K)	+0.23	+0.88	+0.58	+0.43	+1.49	+1.82	+1.54
Peak polar 5° × 5° negative diff (K)	−1.18	−1.53	−1.24	−2.61	−0.85	−1.20	−1.26
Peak 5° × 5° positive diff anywhere (K)	+0.24	+0.88	+0.58	+0.53	+1.49	+1.82	+1.54
Peak 5° × 5° negative diff anywhere (K)	−1.18	−1.53	−1.24	−2.61	−0.85	−1.20	−1.26

been established that the CMIP ensembles are overdispersive for surface air temperature (e.g., Annan and Hargreaves 2010; Yokohata et al. 2012; Bishop and Abramowitz 2013): this narrowing is the result of the EDT accounting for that overdispersiveness. Note that, in most cases, the EDT reduces the growth of ensemble standard deviation throughout the twenty-first century. The ensemble range, shown in the bottom-right and bottom panels of Figs. 6 and 7, respectively, is also consistently reduced.

Figures 8 and 9 show how these changes affect regional temperature projections. To generate these, we first compute the change between 1980–99 average surface temperature and 2080–99 average surface temperature using the multimodel mean in the original ensembles, repeat the process for the dependence-transformed ensembles, and then take the difference of the two. In all scenarios, changes are most marked in higher latitudes, particularly in the Arctic (as much as −2.61°C for RCP8.5 and +1.82°C for SRES A1B). Note that even in cases where there appears little difference at the global scale (RCP6.0 and A2) there may be marked regional differences. Despite the different sign of the affected change at the global scale, there appears to be a consistent change toward Scandinavia, Greenland, and the Barents and Norwegian Seas being cooler than projected by the original CMIP5 ensembles for all RCPs. A similar result, although to a lesser degree, is apparent in the CMIP3 scenarios. The stronger warming trend at the global scale seen for CMIP3 projections is clearly driven by changes at the poles, with consistently increased warming projected in Australia, central Africa, and northwestern South America. Table 2 shows a summary of results. While Figs. 8 and 9 give equal area to each 5° × 5° grid cell, we reiterate that all bias correction, derivation of weights and transformation parameters, and calculation of global averages use area-weighted calculations.

Figure 10 shows the effect of the EDT process on global mean precipitation projections for CMIP5. Projected precipitation increases are reduced at this scale for all RCPs, by −0.003 (−6%), 0.003 (−4%), 0.005 (−5%), and 0.024 mm day^{-1} (−14%) for RCP2.6, RCP4.5, RCP6.0, and RCP8.5, respectively. Growth of ensemble standard deviation is again reduced relative to the original ensemble, although less than in the temperature case. One very interesting feature of Fig. 10 is the discrepancy between observed annual variability from GPCP (shown in blue) and the transformed ensemble standard deviation. Recall that the EDT ensured that observed variance about the CPDF mean estimate (the red line in Fig. 10) was equal to transformed ensemble variance. Recall also that this process was undertaken using all grid cells and monthly time steps of data. Transformation at the per-cell scale apparently has relatively little effect at the global scale. This should serve as a reminder that the process we present is no panacea, and is necessarily constrained by the quality of the model simulations available. We note that issues regarding CMIP models' precipitation variability have been previously documented (e.g., Dai 2006).

Figure 11 shows the regional effects of the EDT on precipitation. These show remarkably consistent patterns, despite the different makeup of contributions to each RCP ensemble. Of particular note is a change to a more El Niño–like mode of precipitation—with a drier western and wetter eastern tropical Pacific—than the default CMIP5 ensembles. The western Pacific warm pool sees reductions in projected rainfall of as much as 410 mm annually while the eastern equatorial Pacific increases by as much as 270 mm. Higher than projected central African, East Asian, and Barents Sea rainfalls are also common features. Results are summarized in Table 3.

6. Discussion and conclusions

Despite the diverse range of changes affected by the ensemble dependence transformation (EDT) of the

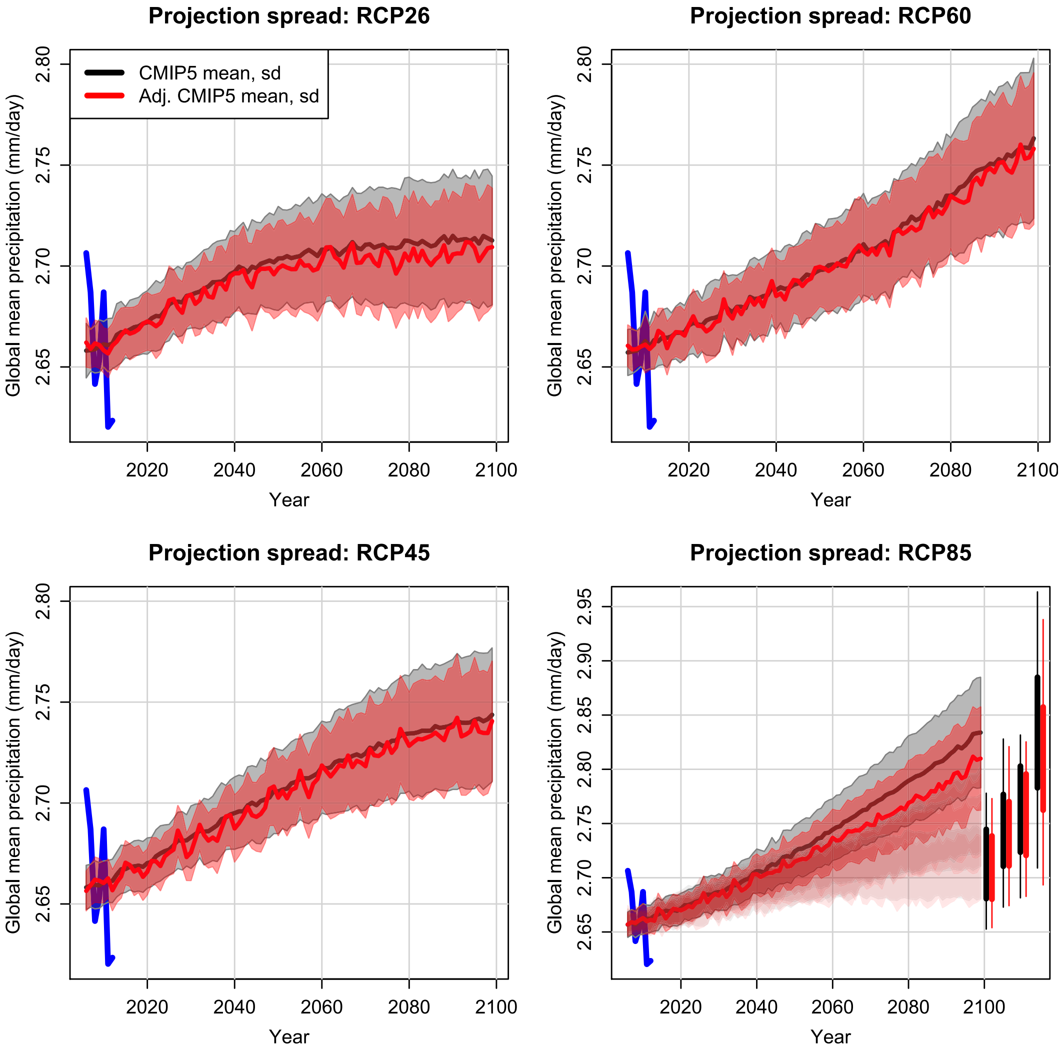

FIG. 10. As in Fig. 6, but for precipitation.

CMIP ensembles, both globally and regionally, we wish to emphasize that these results are not contradictory. The collection of models and the number of simulations from each model submitted to each RCP or emission scenario are very different and are not prescribed by the CMIP experimental protocols. Table 1 shows all simulations used for each RCP in CMIP5 (based on availability at the time of analysis). The differences in contributions to each RCP-based ensemble mean that comparisons between these ensembles reflect not only the differences in forcing conditions prescribed by each RCP but also the differences in the effective model selection strategy for each RCP. Given that the collection of models employed for each scenario is very different, we should expect the EDT to yield different results for different scenarios. Our synthetic model-as-truth experiments suggest that the EDT considerably improves the accuracy of projected CPDFs of temperature and precipitation, giving confidence in the nature of these changes. While not explicitly investigated here, we speculate that the scenario dependence of the EDT corrections may be largely caused by differences in the ensembles' makeup, rather than the forcing conditions.

This possibility should be considered when framing future CMIP protocols. Suppose, for example, a strategy were proposed requiring some institutions to contribute to some scenarios and not others (to lessen the burden of producing simulations). In this case, we would expect increased dependence within each ensemble (fewer models in each scenario) and hence EDT corrections that would probably vary even more starkly between scenarios than they did for the CMIP5 ensembles analyzed here. Such a strategy would make projections for different scenarios more dependent upon ensemble

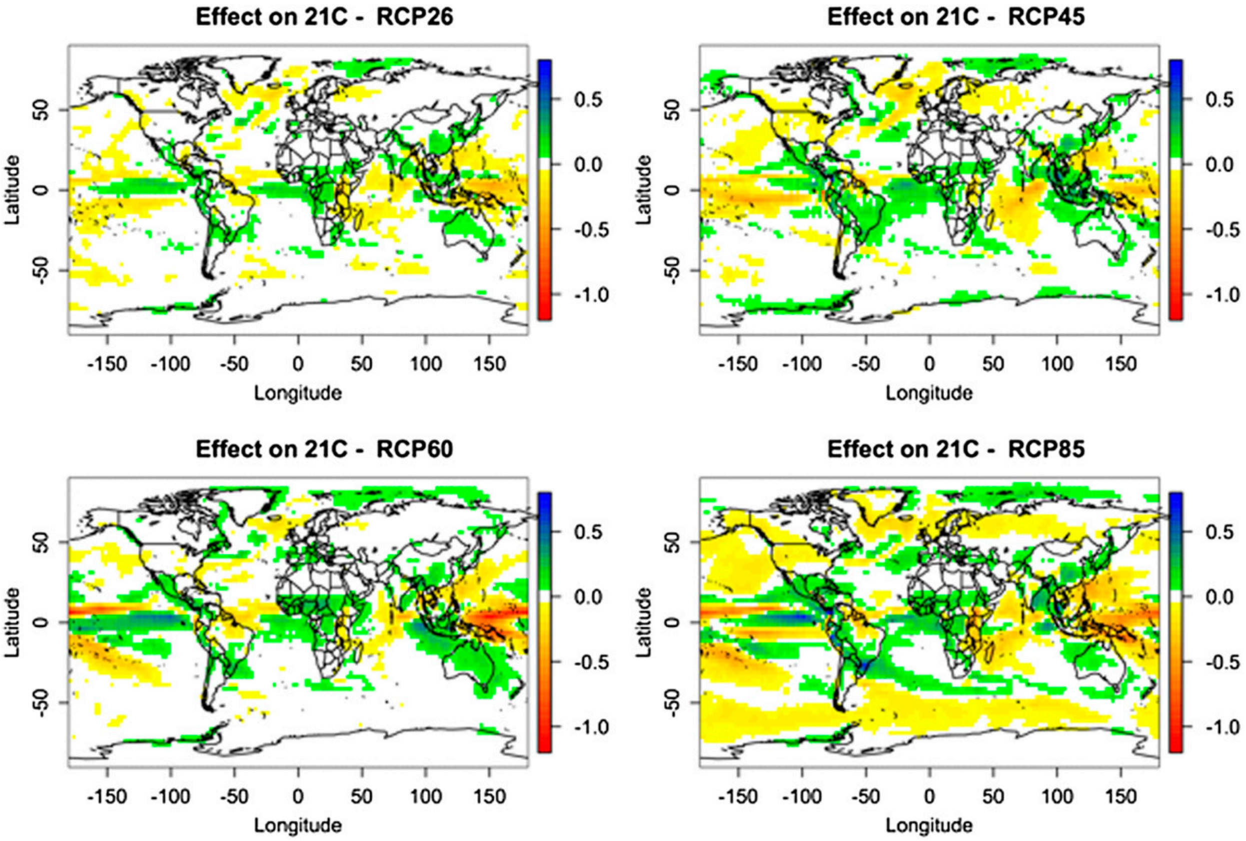

FIG. 11. As in Fig. 8, but for precipitation (mm day^{-1}).

composition and less dependent on the differences in forcing conditions, particularly if a process like the EDT were not used to postprocess the ensembles.

Despite the considerable effort required to produce projections for a range of scenarios, there is a clear need for a consistent sampling strategy across different scenarios that ensures that a diversity of skillful models is included in each emission scenario experiment. Diversity is important because there are two distinct types of uncertainty at play here: uncertainty resulting from natural variability within the system (as defined by the conceptual CPDF) and uncertainty in our ability to create realistic models. Since CMIP ensembles conceptually represent both types of uncertainty, it is perhaps unsurprising that they have been found to be overdispersive (Annan and Hargreaves 2010; Yokohata et al. 2012; Bishop and Abramowitz 2013). However, only an ensemble that is neither underdispersive nor overdispersive in the twentieth century can provide a plausible ensemble projection of the CPDF of the

TABLE 3. The change in various precipitation-related quantities as a result of accounting for model simulation dependence within the CMIP5 ensemble.

	RCP2.6	RCP4.5	RCP6.0	RCP8.5
Mean global precipitation 2005–99 (mm day^{-1})	−0.004	−0.003	−0.002	−0.011
Global precipitation at 2099 (mm day^{-1}) (% increase)	−0.003 (−6%)	−0.003 (−4%)	−0.005 (−5%)	−0.024 (−14%)
Global precipitation std dev (mm day^{-1}) (% increase)	−0.003 (−9%)	−0.003 (−10%)	−0.002 (−5%)	−0.003 (−7%)
Peak 2.5° × 2.5° positive diff anywhere (mm day^{-1})	+0.33	+0.60	+0.50	+0.75
Peak 2.5° × 2.5° negative diff anywhere (mm day^{-1})	−0.49	−0.90	−1.13	−0.81

twenty-first century. In future work we hope to use the EDT to create several such plausible ensembles by subsampling existing CMIP ensembles. The differences in the CPDF estimate from each transformed subsample of ensemble members would then provide an indication of the uncertainty in the CPDF projection associated with model uncertainty. The nature of divergence of CPDF estimates in this case would clearly depend on the composition each subensemble. Subsampling choices might be made to try and maximize the divergence of the CPDF estimates associated with each subsample, or alternatively, one might try and subsample to create independent but equally skillful CPDF estimates. Hopefully, future research will provide further insights into the meaning of such choices and hopefully lead to one or more quantifiably useful strategies.

We also wish to reinforce that the EDT and its underpinnings in the replicate Earth paradigm are something of a shift in the conception of what ensemble spread represents. Spread is typically considered as uncertainty in the estimate of the most likely state (i.e., the mean), deriving from our inability to create a ''perfect model'' (Annan and Hargreaves 2010; Knutti et al. 2010a,b). The EDT focuses on the mean and variance of the CPDF, and so attempts to create an ensemble that replicates the nature of internal system variability. The impact of climate change on societies, biospheres, and the economies is much less a function of the climate mean than it is a function of changes in variability coupled to a change in the mean: for example, droughts, flooding, and heat waves (Alexander and Tebaldi 2012; Seneviratne et al. 2012). A meaningful estimate of the change in CPDF variance over time means that instantaneous estimates of the relative occurrence of unusual events now become more meaningful, since they are now constrained by the EDT. That is, the EDT enables the CMIP ensembles to be turned into CPDF projections whose mean and variance are both plausible.

There are of course many caveats to what we present. We have not spoken at all about differences in the physical mechanisms within models that may lead to the types of corrections we have seen. What we have presented is simply a postprocessing technique that may be useful in circumstances when a best estimate of a particular variable is required. The corrections we made to surface air temperature and precipitation were made independently, and there is no constraint to ensure that they are physically consistent. Nevertheless, we feel confident, particularly after the perfect model experiments detailed in section 4, that the transformed ensemble should provide better estimates of twenty-first-century changes for any particular variable (where sufficient twentieth-century constraining observations are available to derive the transformation). It is also worth reinforcing that the approach assumes that the observational record is long enough to adequately estimate CPDF variance, despite evidence that internal climate variability might operate on much longer time scales (e.g., Ault et al. 2013). This is of course unavoidable, but this potential limitation on the CPDF variance estimate is nevertheless important. We should also restate that this approach obviously cannot create model ability. Applying this approach to a poor ensemble of misleading models will not create a robust projection, and the very real possibility that all or most models are missing key processes remains a key concern. There is always value in improving individual models. Finally, as it stands the EDT relies on an RMSE type metric—essentially constraining mean and spatiotemporal variability—when in many applications other metrics are more important (e.g., extreme values).

In conclusion, we have found that the EDT has a considerable effect on projected surface air temperature and precipitation for the twenty-first century. Our confidence in the nature of the changes effected by the EDT is strengthened by perfect model experiments where performance in the out-of-sample twenty-first century is comparable to the in-sample twentieth century. Specifically, RMSD between the EDT ensemble mean and ''observations'' is on average 30% less than the untransformed ensemble in the (out of sample) twenty-first century. In addition, the variance of the EDT ensemble matches the variance of ''observations'' about the ensemble mean much better than in the untransformed ensemble. In some of the real-world CMIP scenarios, we found that the EDT changed projected global warming by around 15%, with regional differences exceeding 2.5°C, depending on the composition of the ensemble. Projected precipitation changes effected by accounting for dependence are similar, with changes as large as 14% of the projected global change and regional changes as large as 410 mm annually over a two-decade period. Finally, the EDT offers a clear theoretical interpretation of ensemble spread as a quantification of internal climate system variability.

Acknowledgments. We acknowledge the support of the Australian Research Council Centre of Excellence for Climate System Science (CE110001028). CHB gratefully acknowledges support from the U.S. Office of Naval Research Grant N0001413WX00008. The GPCP combined precipitation data were developed and computed by the NASA Goddard Space Flight Center Laboratory for Atmospheres as a contribution to the GEWEX Global Precipitation Climatology Project.

GPCP data were provided by the NOAA/OAR/ESRL/PSD, Boulder, Colorado, from their website (http://www.esrl.noaa.gov/psd/). All analysis and transformation code are available from the corresponding author upon request.

REFERENCES

Abramowitz, G., 2010: Model independence in multi-model ensemble prediction. *Aust. Meteor. Ocean. J.,* **59,** 3–6.

Adler, R. F., and Coauthors, 2003: The version-2 Global Precipitation Climatology Project (GPCP) monthly precipitation analysis (1979–present). *J. Hydrometeor.,* **4,** 1147–1167, doi:10.1175/1525-7541(2003)004<1147:TVGPCP>2.0.CO;2.

Alexander, L., and C. Tebaldi, 2012: Climate and weather extremes: Observations, modelling, and projections. *The Future of the World's Climate,* A. Henderson-Sellers and K. McGuffie, Eds., Elsevier Science, 253–288.

Annan, J. D., and J. C. Hargreaves, 2010: Reliability of the CMIP3 ensemble. *Geophys. Res. Lett.,* **37,** L02703, doi:10.1029/2009GL041994.

Ault, T. R., J. E. Cole, J. T. Overpeck, G. T. Pederson, S. St. George, B. Otto-Bliesner, C. A. Woodhouse, and C. Deser, 2013: The continuum of hydroclimate variability in western North America during the last millennium. *J. Climate,* **26,** 5863–5878, doi:10.1175/JCLI-D-11-00732.1.

Bishop, C., and G. Abramowitz, 2013: Climate model dependence and the replicate Earth paradigm. *Climate Dyn.,* **41** (3–4), 885–900, doi:10.1007/s00382-012-1610-y.

Brohan, P., J. J. Kennedy, I. Harris, S. F. B. Tett, and P. D. Jones, 2006: Uncertainty estimates in regional and global observed temperature changes: A new dataset from 1850. *J. Geophys. Res.,* **111,** D12106, doi:10.1029/2005JD006548.

CMIP, cited 2013a: IPCC/CMIP5 AR5 timetable. [Available online at http://cmip-pcmdi.llnl.gov/cmip5/ipcc_ar5_timetable.html.]

——, cited 2013b: CMIP5—Data access—Availability. [Available online at http://cmip-pcmdi.llnl.gov/cmip5/availability.html.]

Collins, M., S. F. B. Tett, and C. Cooper, 2001: The internal climate variability of HadCM3, a version of the Hadley Centre coupled model without flux adjustments. *Climate Dyn.,* **17,** 61–81, doi:10.1007/s003820000094.

Dai, A., 2006: Precipitation characteristics in eighteen coupled climate models. *J. Climate,* **19,** 4605–4630, doi:10.1175/JCLI3884.1.

Deser, C., A. Phillips, V. Bourdette, and H. Teng, 2012: Uncertainty in climate change projections: The role of internal variability. *Climate Dyn.,* **38** (3–4), 527–546, doi:10.1007/s00382-010-0977-x.

Ehret, U., E. Zehe, V. Wulfmeyer, K. Warrach-Sagi, and J. Liebert, 2012: Should we apply bias correction to global and regional climate model data? *Hydrol. Earth Syst. Sci.,* **16,** 3391–3404, doi:10.5194/hess-16-3391-2012.

England, M. H., and Coauthors, 2014: Recent intensification of wind-driven circulation in the Pacific and the ongoing warming hiatus. *Nat. Climate Change,* **4,** 222–227, doi:10.1038/nclimate2106.

Gneiting, T., and A. E. Raftery, 2005: Weather forecasting with ensemble methods. *Science,* **310,** 248–249, doi:10.1126/science.1115255.

Hamill, T. M., S. L. Mullen, C. Snyder, D. P. Baumhefner, and Z. Toth, 2000: Ensemble forecasting in the short to medium range: Report from a workshop. *Bull. Amer. Meteor. Soc.,* **81,** 2653–2664, doi:10.1175/1520-0477(2000)081<2653:EFITST>2.3.CO;2.

Haughton, N., G. Abramowitz, A. Pitman and S. Phipps, 2014: On the generation of climate model ensembles. *Climate Dyn.,* **43,** 2297–2308, doi:10.1007/s00382-014-2054-3.

James, I. N., and P. M. James, 1989: Ultra low frequency variability in a simple global circulation model. *Nature,* **342,** 53–55, doi:10.1038/342053a0.

Jun, M., R. Knutti, and D. Nychka, 2008: Spatial analysis to quantify numerical model bias and dependence: How many climate models are there? *J. Amer. Stat. Assoc.,* **103,** 934–947, doi:10.1198/016214507000001265.

Knutti, R., G. Abramowitz, M. Collins, V. Eyring, P. J. Gleckler, B. Hewitson, and L. Mearns, 2010a: Good practice guidance paper on assessing and combining multi model climate projections. IPCC Expert Meeting on Assessing and Combining Multi Model Climate Projections Meeting Rep., 15 pp. [Available online at http://www.ipcc-wg2.gov/meetings/EMs/IPCC_EM_MME_GoodPracticeGuidancePaper.pdf.]

——, R. Furrer, C. Tebaldi, J. Cermak, and G. A. Meehl, 2010b: Challenges in combining projections from multiple models. *J. Climate,* **23,** 2739–2758, doi:10.1175/2009JCLI3361.1.

Macadam, I., A. J. Pitman, P. H. Whetton, and G. Abramowitz, 2010: Ranking climate models by performance using actual values and anomalies: Implications for climate change impact assessments. *Geophys. Res. Lett.,* **37,** L16704, doi:10.1029/2010GL043877.

Meehl, G. A., C. Covey, T. Delworth, M. Latif, B. McAvaney, J. F. B. Mitchell, R. J. Stouffer, and K. E. Taylor, 2007a: The WCRP CMIP3 multimodel data set: A new era in climate change research. *Bull. Amer. Meteor. Soc.,* **88,** 1383–1394, doi:10.1175/BAMS-88-9-1383.

——, and Coauthors, 2007b: Global climate projections. *Climate Change 2007: The Physical Science Basis,* S. Solomon et al., Eds., Cambridge University Press, 747–845.

Meinshausen, M., and Coauthors, 2011: The RCP greenhouse gas concentrations and their extension from 1765 to 2300. *Climatic Change,* **109** (1–2), 213–241, doi:10.1007/s10584-011-0156-z.

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 dataset. *J. Geophys. Res.,* **117,** D08101, doi:10.1029/2011JD017187.

Nakicenovic, N., and R. Swart, Eds., 2000: *Special Report on Emissions Scenarios.* Cambridge University Press, 599 pp.

Potempski, S., and S. Galmarini, 2009: *Est modus in rebus*: Analytical properties of multi-model ensembles. *Atmos. Chem. Phys.,* **9,** 9471–9489, doi:10.5194/acp-9-9471-2009.

Seneviratne, S. I., and Coauthors, 2012: Changes in climate extremes and their impacts on the natural physical environment. *Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation,* C. B. Field et al., Eds., Cambridge University Press, 109–230.

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, doi:10.1175/BAMS-D-11-00094.1.

Tebaldi, C., and R. Knutti, 2007: The use of the multimodel ensemble in probabilistic climate projections. *Philos. Trans. Roy. Soc.,* **365A,** 2053–2075, doi:10.1098/rsta.2007.2076.

Yokohata, T., and Coauthors, 2012: Reliability of multi-model and structurally different single-model ensembles. *Climate Dyn.,* **39** (3–4), 599–616, doi:10.1007/s00382-011-1203-1.