Emergent Constraints on the Large-Scale Atmospheric Circulation and Regional Hydroclimate: Do They Still Work in CMIP6 and How Much Can They Actually Constrain the Future?

ISLA R. SIMPSON,[a] KAREN A. MCKINNON,[b] FRANCES V. DAVENPORT,[c] MARTIN TINGLEY,[d] FLAVIO LEHNER,[a,e] ABDULLAH AL FAHAD,[f] AND DI CHEN[g]

[a] *Climate and Global Dynamics Laboratory, National Center for Atmospheric Research, Boulder, Colorado*
[b] *Department of Statistics, Institute of the Environment, University of California, Los Angeles, Los Angeles, California*
[c] *Department of Earth System Science, Stanford University, Stanford, California*
[d] *Boulder, Colorado*
[e] *Department of Earth and Atmospheric Sciences, Cornell University, Ithaca, New York*
[f] *George Mason University, Fairfax, Virginia*
[g] *Atmospheric and Oceanic Sciences, University of California, Los Angeles, California*

(Manuscript received 19 January 2021, in final form 29 April 2021)

ABSTRACT: An emergent constraint (EC) is a statistical relationship, across a model ensemble, between a measurable aspect of the present-day climate (the predictor) and an aspect of future projected climate change (the predictand). If such a relationship is robust and understood, it may provide constrained projections for the real world. Here, models from phase 6 of the Coupled Model Intercomparison Project (CMIP6) are used to revisit several ECs that were proposed in prior model intercomparisons with two aims: 1) to assess whether these ECs survive the partial out-of-sample test of CMIP6 and 2) to more rigorously quantify the constrained projected change than previous studies. To achieve the latter, methods are proposed whereby uncertainties can be appropriately accounted for, including the influence of internal variability, uncertainty on the linear relationship, and the uncertainty associated with model structural differences, aside from those described by the EC. Both least squares regression and a Bayesian hierarchical model are used. Three ECs are assessed: (i) the relationship between Southern Hemisphere jet latitude and projected jet shift, which is found to be a robust and quantitatively useful constraint on future projections; (ii) the relationship between stationary wave amplitude in the Pacific–North American sector and meridional wind changes over North America (with extensions to hydroclimate), which is found to be robust but improvements in the predictor in CMIP6 result in it no longer substantially constraining projected change in either circulation or hydroclimate; and (iii) the relationship between ENSO teleconnections to California and California precipitation change, which does not appear to be robust when using historical ENSO teleconnections as the predictor.

KEYWORDS: Atmosphere; Stationary waves; Jets; Precipitation; Climate models

1. Introduction

As we grapple to predict the future of the climate system, there are three sources of uncertainty we must contend with: scenario uncertainty, internal variability, and model response uncertainty (Hawkins and Sutton 2009; Lehner et al. 2020). Scenario uncertainty arises because we do not know exactly how anthropogenic forcings will evolve in the future and is dealt with by considering a range of future forcing scenarios that span the range of possible societal outcomes (e.g., O'Neill et al. 2013). Internal variability arises because the Earth system internally generates its own variability and, as our singular climate system evolves, the climate state we experience will be a combined result of both anthropogenically forced change and this internal variability (e.g., Deser et al. 2012). For the most part, this source of uncertainty is irreducible, but can nevertheless be quantified; it is a certain uncertainty (Deser 2020). Finally, model response uncertainty arises because we are attempting to predict the future with imperfect models. While model development and enhanced computing capabilities are continually aimed at reducing this uncertainty, we must, at the same time, come up with creative ways of either interpreting model response uncertainty (Shepherd et al. 2018) or reducing it (Hall et al. 2019).

"Emergent constraints" (ECs) are a potential way to reduce model response uncertainty (Hall et al. 2019; Brient 2020). These are statistical relationships between the modeled representation of a measurable aspect of the present-day climate (the predictor) and some aspect of future projected change (the predictand), with the expectation that the predictor and predictand are linked somehow in a physically meaningful way. Most commonly, these constraints "emerge" from multimodel ensembles such as the Coupled Model Intercomparison Project (CMIP; Taylor et al. 2012; Eyring et al. 2016) but they can also be found, or tested, in perturbed physics ensembles with an individual model (Kamae et al. 2016; Wagman and Jackson 2018).

Since the potential of ECs was brought to the fore by Hall and Qu (2006) in their analysis of snow-albedo feedbacks, they have been applied to climate sensitivity and cloud feedbacks (Caldwell et al. 2018, and references therein), carbon

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/JCLI-D-21-0055.s1.

Corresponding author: Isla R. Simpson, islas@ucar.edu

DOI: 10.1175/JCLI-D-21-0055.1

cycle feedbacks (Cox et al. 2013; Wenzel et al. 2014), ocean productivity (Kwiatkowski et al. 2017), sea ice loss (Boé et al. 2009; Massonnet et al. 2012), and aspects of the large-scale circulation and hydroclimate (Kidston and Gerber 2010; O'Gorman 2012; Simpson and Polvani 2016; Simpson et al. 2016; Li et al. 2017; Lehner et al. 2019; Chen et al. 2020).

As embodied in the views of two recent perspective articles (Hall et al. 2019; Brient 2020), great care must be taken in assessing the validity and usefulness of an EC, since spurious significant relationships can be found within multimodel ensembles, purely by chance (Caldwell et al. 2014). Hall et al. (2019) put forth a framework whereby a proposed EC, based on a strong statistical relationship, can become verified. This involves accompanying the EC with a plausible physical mechanism, verifying that this mechanism is at work in the model ensemble, and assessing whether the EC survives out-of-sample testing. Ensuring that a relationship survives out-of-sample testing using another ensemble or, alternatively, using dedicated sensitivity experiments within a single model (e.g., van Niekerk et al. 2017) helps to establish that the relationship is robust and is indeed indicative of a true underlying physical relationship between predictor and predictand. Brient (2020) also discusses the importance of accounting for the many uncertainties involved when quantifying constrained future projections.

Prior EC studies address this issue of the quantification of future projected change with varying degrees of rigor. Many studies stop short of providing a quantitative estimate for the future and, instead, simply point out the constraining relationship and discuss how the observed predictor compares to the model distribution to draw qualitative conclusions (Hall and Qu 2006; Trenberth and Fasullo 2010; Fasullo and Trenberth 2012; Su et al. 2014; Tian 2015; Simpson and Polvani 2016; Lipat et al. 2017). Others are more quantitative by using a linear regression combined with the observed value of the predictor to project the future, but often without, or only a partial, consideration of the uncertainties involved (Volodin 2008; Sherwood et al. 2014). Some do account for uncertainty in the regression coefficients (Huber et al. 2010; O'Gorman 2012; Simpson et al. 2016) but neglect other potential sources of model spread that are not described by the EC, while others quantify the uncertainty based on the residuals from the linear regression fit, which may be a more encompassing approach (Bracegirdle and Stephenson 2013; Cox et al. 2013), although Bowman et al. (2018) highlight the importance of also incorporating observational uncertainty. Brient (2020) provides a clear example where failure to adequately account for uncertainties can lead to overly constrained projections and, instead, opts to use a model weighting procedure to provide a constrained distribution of future projected change—an approach that has been implemented by a number of other studies with varying degrees of sophistication (Hargreaves et al. 2012; Massonnet et al. 2012; Zhai et al. 2015).

Here, we revisit three ECs that relate to the large-scale atmospheric circulation and regional hydroclimate with two primary goals. The first is to assess whether these constraints, which were previously found in CMIP5 models, still exist in CMIP6. While CMIP6 compared to CMIP5 may not be a completely out-of-sample test, passing this test will improve confidence that the EC is real and move it along the path toward the confirmed category (Hall et al. 2019). Our second goal is to more rigorously quantify the extent to which these constraints can actually constrain future projections, by proposing a new approach to adequately incorporate the variety of uncertainties that are involved in the calculation. This approach involves linear regression, using either least squares regression or a Bayesian hierarchical model, combined with sampling from large ensembles from multiple models, to quantify the constrained future change via the EC.

The three ECs to be assessed are 1) the relationship between the climatological latitude and future projected poleward shift of the Southern Hemisphere (SH) jet stream (Kidston and Gerber 2010; Simpson and Polvani 2016), hereafter referred to as the SHJET constraint; 2) the relationship between the climatological amplitude of intermediate scale stationary waves in the Pacific–North American sector and future projected meridional wind change over North America, with extension to North American hydroclimate (Simpson et al. 2016), hereafter referred to as the VWIND constraint; and 3) the relationship between a model's representation of El Niño–Southern Oscillation (ENSO) teleconnections to California precipitation and future projected California precipitation change (Allen and Luptowitz 2017), hereafter referred to as the CALP constraint.

Section 2 describes the model and observation-based datasets used and methods are outlined in section 3. The SHJET, VWIND, and CALP constraints are then assessed in sections 4, 5, and 6, respectively. Discussion is provided in section 7 followed by conclusions in section 8.

2. Model and observation-based data

Monthly zonal wind (U), meridional wind (V), precipitation (pr), surface temperature (T_s) and surface air temperature (T_{2m}) data from the CMIP5 and CMIP6 models summarized in Table 1 are used, after interpolating fields to a common 1° grid using bilinear interpolation. For the SHJET and VWIND constraint, we take the "past" to be the period 1979–2014 under the "historical" forcing scenario (CMIP5 historical members are combined with the corresponding RCP8.5 member to extend the historical period out to 2014). We take the "future" to be 2070–99 of the RCP8.5 scenario in CMIP5 (Meinshausen et al. 2011; Lamarque et al. 2011) and the SSP5-8.5 scenario in CMIP6 (Kriegler et al. 2017)—that is, using the forcing scenario in each CMIP that reaches ~8.5 W m^{-2} radiative forcing by the end of the century. For the CALP constraint, the linear trend in precipitation between 2006 and 2099 is considered, while both the 2006–99 and 1948–2014 periods are used to assess the representation of ENSO teleconnections.

To quantify the influence of internal variability on uncertainty in both the predictor and predictand, we use initial condition large ensembles (LEs) from five models run under the RCP8.5 scenario: CanESM2, 50 members; CESM1-CAM5, 40 members; CSIRO-Mk3.6.0, 30 members; GFDL-CM3, 20 members; and MPI-ESM, 100 members (Deser et al. 2020).

For observed U and V we use ERA5 (Hersbach et al. 2020), ERA-Interim (Dee et al. 2011), MERRA2 (Gelaro et al. 2017),

TABLE 1. A summary of the CMIP5 and CMIP6 models and number of ensemble members used; hist refers to the historical simulations, 8.5 refers to the forcing scenario that results in 8.5 W m^{-2} radiative forcing by the end of the century (RCP8.5 for CMIP5 and SSP5-8.5 for CMIP6). Superscripts from 1 to 22 indicate the 22 CMIP5 and CMIP6 models that are considered to be predecessors and successors to test the independence of CMIP5 and CMIP6. (Expansions of most model names are available online at http://www.ametsoc.org/PubsAcronymList.)

CMIP5			CMIP6		
Name	hist	8.5	Name	hist	8.5
ACCESS1.0	1	1	ACCESS-CM2[1]	2	1
ACCESS1.3[1]	1	1	ACCESS-ESM1-5	3	3
BCC-CSM1.1	1	1	AWI-CM-1-1-MR	5	1
BCC-CSM1.1-m[2]	1	1	BCC-CSM2-MR[2]	3	1
BNU-ESM	1	1	CAMS-CSM1-0	1	2
CanESM2[3]	5	5	CanESM5[3]	25	25
CCSM4	6	6	CanESM5-CanOE	3	3
CESM1-BGC	1	1	CESM2[4]	10	2
CESM1-CAM5[4]	3	3	CESM2-WACCM[5]	3	1
CESM1-WACCM[5]	1	1	CIESM	3	1
CMCC-CM	1	1	CMCC-CM2-SR5[6]	1	1
CMCC-CMS[6]	1	1	CNRM-CM6-1[7]	15	6
CNRM-CM5[7]	5	5	CNRM-CM6-1-HR	1	1
CSIRO-Mk3.6.0	10	10	CNRM-ESM2-1	5	5
EC-EARTH[8]	1	1	EC-Earth3[8]	10	7
FGOALS-g2[9]	1	1	EC-Earth3-Veg	4	3
FIO-ESM[10]	3	3	FGOALS-f3-L	3	1
GFDL-CM3[11]	1	1	FGOALS-g3[9]	3	1
GFDL-ESM2G	1	1	FIO-ESM-2-0[10]	3	3
GFDL-ESM2M[12]	1	1	GFDL-CM4[11]	1	1
GISS-E2-H	2	2	GFDL-ESM4[12]	1	1
GISS-E2-R[13]	2	2	GISS-E2-1-G[13]	10	1
HadGEM2-AO	1	1	HadGEM3-GC31-LL[14]	4	3
HadGEM2-CC	3	3	HadGEM3-GC31-MM	4	3
HadGEM2-ES[14]	3	3	INM-CM4-8	1	1
INM-CM4[15]	1	1	INM-CM5-0[15]	8	1
IPSL-CM5A-LR[16]	4	4	IPSL-CM6A-LR[16]	32	1
IPSL-CM5A-MR	1	1	KACE-1-0-G	3	1
IPSL-CM5B-LR	1	1	MCM-UA-1-0	1	1
MIROC5[17]	3	3	MIROC6[17]	10	3
MIROC-ESM[18]	1	1	MIROC-ES2L[18]	3	1
MIROC-ESM-CHEM	1	1	MPI-ESM1-2-HR[20]	10	1
MPI-ESM-LR[19]	3	3	MPI-ESM1-2-LR[19]	10	1
MPI-ESM-MR[20]	1	1	MRI-ESM2-0[21]	5	1
MRI-CGCM3[21]	1	1	NESM3	5	2
NorESM1-M	1	1	NorESM2-LM[22]	3	1
NorESM1-ME[22]	1	1	NorESM2-MM	1	1
			UKESM1-0-LL	4	5

and JRA-55 (Kobayashi et al. 2015) reanalyses. For T_s over ocean we use HadISST (Rayner et al. 2003), ERSSTv3b (Smith et al. 2008), and ERSSTv5 (Huang et al. 2017), and for pr we use CRUTS (Harris et al. 2014) and GPCC (Schneider 2018) station-based observations.

3. Emergent constraints methodology

The essence of an EC is a relationship (typically linear) between a model's representation of a present-day quantity (x) and its projected future change (Δ) in a quantity (y). Using a linear fit:

$$\Delta y(i) = \alpha + \beta x(i) + \epsilon(i), \quad (1)$$

where α and β are the regression coefficients, $\varepsilon(i)$ are the residuals of the fit, and i refers to individual model data points ($i = 1, \ldots, N$) after averaging over the ensemble members available (see the schematic example in Fig. 1a). The measured $(x, \Delta y)$ for a given model may differ from the true $(\overline{x}, \Delta\overline{y})$ due to internal variability given the limited ensemble size for each model. We will use $\overline{(\,.\,)}$ throughout to refer to the "true" values, absent internal variability. The residuals (ε) will have a component that results from internal variability ($\varepsilon_{\mathrm{IV}}$) and a component that arises from other intermodel differences in $\Delta\overline{y}$ that are not described by the emergent constraint, which we will refer to as δ. If the effects of internal variability could be neglected, then

$$\Delta\overline{y}(i) = \alpha + \beta\overline{x}(i) + \delta(i). \quad (2)$$

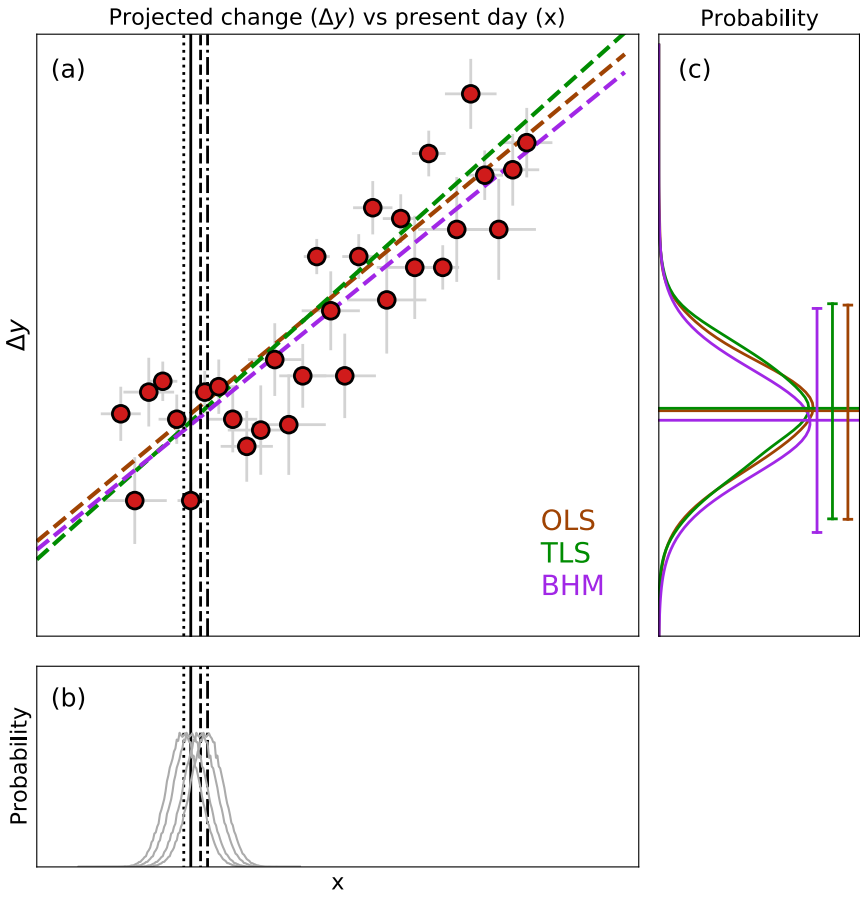

FIG. 1. Illustrative depiction of the EC method using synthetic data. (a) The relationship between the projected change (Δy) and the present day climatology (x) for the models depicted by the red points. Gray crosshairs depict the uncertainty on each red point by the $\pm 1.96\sigma$ range. Black vertical lines show four different measurements of x_E while the brown, green, and purple dashed lines show the best fitting linear regression line using the OLS, TLS, and BHM methods. (b) The distribution of possible true values of the real world climatology $\overline{x}_E$ given by PDFs that reflect the uncertainty due to internal variability, centered on each observed value. (c) The probability distributions of the real world change along with its mean (horizontal lines) and the 95% confidence interval (vertical ranges) for each method. This PDF considers the uncertainty in the best fitting regression line, the uncertainty in the true real world value of x, and the potential influence of internal variability and other aspects of the forced response not explained by the EC, using the method outlined in section 3.

The variance in modeled Δy values, $\sigma^2(\Delta y)$, can be partitioned into a component that is explained by the EC (σ^2_{EC}) and the remainder (σ^2_ϵ). The term σ^2_ϵ then consists of components due to internal variability (σ^2_{IV}) and other intermodel differences that are not explained by the EC (σ^2_δ); that is,

$$\sigma^2(\Delta y) = \sigma^2_{\mathrm{EC}} + \sigma^2_{\mathrm{IV}} + \sigma^2_\delta, \tag{3}$$

where we have assumed ε, δ, and $\varepsilon_{\mathrm{IV}}$ are normally distributed such that $\epsilon \sim \mathcal{N}(0, \sigma^2_\epsilon)$, $\delta \sim \mathcal{N}(0, \sigma^2_\delta)$ and $\epsilon_{\mathrm{IV}} \sim \mathcal{N}(0, \sigma^2_{\mathrm{IV}})$. Combining the EC (1) with the observed value of x for the real world [x_{E}, with $(.)_{\mathrm{E}}$ referring to Earth], the future change for the real world (Δy_E) can be predicted via

$$\Delta y_E = \alpha + \beta x_E + \epsilon_{\mathrm{IV}} + \delta. \tag{4}$$

The first two terms on the right refer to the component predicted by the EC, the third term represents the Δy that could arise due to internal variability in one realization, and the final term refers to the other contributions to the forced change in the real world that are not explained by the EC.

Each component on the right of (4) is uncertain. With only a finite number of models, with finite ensemble sizes, α and β are not known exactly; also, x_{E} may deviate from $\overline{x}_E$ due to observational error and internal variability (Fig. 1b), $\varepsilon_{\mathrm{IV}}$ is an irreducible uncertainty, and we may not know the role of other forced responses in the real world, not described by the EC (δ), although this uncertainty has the potential to be reduced through additional emergent constraints as they are discovered. A further assumption is made that the process representation within the models is close enough to that of the real

world, that the real world $\Delta\overline{y}$ will depend on $\overline{x}$ in a similar way (Williamson and Sansom 2019).

While we cannot know the true values of any of the uncertain parameters, we can model them as probability distribution functions (PDFs), based on the information we have, to determine a PDF for Δy_E (depicted schematically in Fig. 1c), which reflects the added information from the EC. The idea is then that this probability distribution will be more constrained than what would be derived from the model ensemble directly. For each constraint, distributions for Δy_E will be derived using two different types of least squares regression and a Bayesian hierarchical model, now described.

a. *Ordinary and total least squares regressions*

For both ordinary and total least squares regressions (OLS and TLS), α and β are determined by minimizing a loss function that depends on the squared residuals and our estimate of internal variability in the form of the standard deviation (σ_x or $\sigma_{\Delta y}$, which will depend on model ensemble size) and assuming that $\overline{x}$ and $\Delta\overline{y}$ are represented by normal distributions with these standard deviations centered on x and Δy. Section 3c below describes how σ_x and $\sigma_{\Delta y}$ are estimated. OLS and TLS only differ in the method used to derive α and β. For OLS, all errors are assumed to be in the dependent variable Δy. A weighted approach allows for a different $\sigma_{\Delta y}$ for each model by finding the α and β that minimize $\sum_{i=1}^{N}\{[\Delta y(i)-\alpha-\beta x(i)]/\sigma_{\Delta y}(i)\}^2$. For TLS, errors in the x direction (σ_x) are also accounted for by instead minimizing $\sum_{i=1}^{N}[\Delta y(i)-\alpha-\beta x(i)]^2/[\sigma_{\Delta y}(i)^2+\beta^2\sigma_x(i)^2]$.

To decompose the intermodel variance according to (3), $\sigma_{EC}^2=\sigma^2(\Delta y)-\sigma_\epsilon^2$ since the variance in Δy consists of variance explained by the EC ($\sigma_{\rm EC}^2$) plus the variance of the residuals (σ_ϵ^2). The internal variability component $\sigma_{\rm IV}^2$ is estimated via $\sigma_{\rm IV}^2=\beta^2\sigma_x^2+\sigma_{\Delta y}^2$ [1] and σ_δ^2 is then estimated as the remainder of σ_ϵ^2. While this procedure does not guarantee that σ_δ^2 is positive, as it needs to be, we find that $\sigma_{\rm IV}^2$ is less than σ_ϵ^2 in all cases.

To provide a constrained future projection for the real world, PDFs for each of the parameters on the right-hand side of (4) are constructed. A PDF of 1000 α, β combinations is determined by bootstrapping N models from the N available, with replacement, and recalculating the regression fit. A PDF of $\overline{x}_E$ is estimated using a variety of observational products as described in section 2 (to account for observational error) and then modeling $\overline{x}_E$ as a normal distribution centered on each observed value with standard deviation σ_x (to account for the uncertainty due to internal variability); that is, we assume that the observed x_E is the most likely true value $\overline{x}_E$, an assumption that, of course, cannot be tested in the context of the single observational record. Internal variability will also play a role in the future − past difference, so to account for this combined with the other forced contributions that are unrelated to the EC ($\Delta y_{\rm IV}+\delta_{\rm E}$) 1000 values are sampled from a normal distribution with variance $\sigma_\epsilon^2-\beta^2\sigma_x^2$, which is equivalent to $\sigma_{\Delta y}^2+\sigma_\delta^2$ (see footnote 1). Combining all permutations of these samples gives 1 billion values of Δy_E according to (4) that represent our constrained distribution for the real world (Fig. 1c).

The real world forced response ($\Delta\overline{y}_E$), absent internal variability, can be estimated by applying a similar procedure but without the internal variability component, that is, $\Delta\overline{y}_E=\alpha+\beta x_E+\delta_E$. To sample δ_E we estimate the variance of δ by $\sigma_\delta^2=\sigma_\epsilon^2-\sigma_{\rm IV}^2$ and then 1000 values of δ_E are sampled from a normal distribution with this variance.

b. *The Bayesian hierarchical model*

The Bayesian hierarchical model (BHM) fits the regression model (2) by modeling the "true" $\overline{x}$ and $\Delta\overline{y}$ (uncontaminated by internal variability) as probability distributions based on x, Δy, σ_x, and $\sigma_{\Delta y}$, and the correlation between the internal variability uncertainties ($r_{x\Delta y}$).

The BHM is described in more detail in the appendix, so here we summarize how its output is used to decompose the variance in Δy via (3) and to estimate the constrained distribution of Δy_E. A product of the BHM is 1000 estimates of (α, β, σ_δ^2, δ_x^2), where δ_x^2 is the standard deviation of $\overline{x}$ (i.e., the spread across the "true" values of the predictor in the climate models). For each of these 1000 estimates, the variance explained by the EC is given by $\sigma_{\rm EC}^2=\beta^2\delta_x^2$, the variance explained by intermodel differences in the forced response not described by the EC is given by σ_δ^2, and the variance explained by internal variability is given by $\sigma_{\Delta y}^2(1-r_{x\Delta y}^2)$. The fraction of variance explained by each component is estimated from each BHM sample separately and then the mean over the 1000 estimates is displayed.

The BHM provides a ready formalism for sampling the various uncertainties already described for OLS and TLS. To quantify the constrained distribution for the forced response, the 1000 (α, β, σ_δ^2) combinations are combined with 1000 estimates of $\overline{x}_E$ (sampled in the same way as OLS and TLS). The constrained distribution of the forced response $\Delta\overline{y}$ is given by $\alpha+\beta\overline{x}_E+\mathcal{N}(0,\sigma_\delta^2)$, where $\mathcal{N}(0,\sigma_\delta^2)$ represents a random sample from a normal distribution with zero mean and variance σ_δ^2. Combining the 1000 estimates of (α, β, σ_δ^2) with 1000 estimates of $\overline{x}_E$ and 1000 samples from $\mathcal{N}(0,\sigma_\delta^2)$ gives 1 billion values of Δy to form the constrained distribution.

To construct the constrained distribution for the forced response plus internal variability, Eq. (A10) is used, where each of the 1 billion Δy values derived above are combined with $r_{x\Delta y}(\sigma_{\Delta y}/\sigma_x)(x_E-\overline{x}_E)$ and a random sample from a normal distribution with variance equal to $\sigma_{\Delta y}^2(1-r_{x\Delta y}^2)$, where $\overline{x}_E$ is the actual observed value and x_E are the 1000 estimates sampled from the PDF, corresponding to an assumption that the most likely value of $\overline{x}_E$ is that which we have observed.

c. *Estimating σ_x, $\sigma_{\Delta y}$, and $r_{x\Delta y}$*

The above procedures rely on estimates of the uncertainty due to internal variability on x and Δy as represented by σ_x and

[1] This can be shown by considering $\Delta\overline{y}$ and $\overline{x}$ to deviate from the measured Δy and x by amounts $\varepsilon_{\Delta y}$ and ε_x due to internal variability [i.e., $\Delta y(i)+\varepsilon_{\Delta y}=\alpha+\beta[x(i)+\varepsilon_x]+\delta(i)$]. Rearranging gives $\Delta y(i)-\alpha-\beta x(i)=\varepsilon(i)=\beta\varepsilon_x-\varepsilon_{\Delta y}+\delta(i)$, so σ_ϵ^2 can be partitioned according to $\sigma_\epsilon^2=\beta^2\sigma_x^2+\sigma_{\Delta y}^2+\sigma_\delta^2$, where $\beta^2\sigma_x^2+\sigma_{\Delta y}^2$ represents the contribution due to internal variability ($\sigma_{\rm IV}^2$).

$\sigma_{\Delta y}$. For OLS and TLS, it is assumed that the PDFs of $\overline{x}$ and $\Delta\overline{y}$ are Gaussian and centered on x and Δy, with standard deviations σ_x and $\sigma_{\Delta y}$, respectively. For BHM, the correlation between these uncertainties due to internal variability ($r_{x\Delta y}$) is also incorporated by assuming a bivariate normal distribution centered on x and Δy [Eq. (A8)].

Therefore, we need to estimate values of σ_x, $\sigma_{\Delta y}$, and $r_{x\Delta y}$ and this must necessarily account for the number of ensemble members available for a given model since the more ensemble members there are, the smaller σ_x and $\sigma_{\Delta y}$ will be. Ideally, we would also want to account for the fact that different models may have different representations of internal variability and, therefore, different values of σ_x and $\sigma_{\Delta y}$. But, as will be described individually for the constraints discussed below, attempts at quantification of σ_x and $\sigma_{\Delta y}$ for models with a small number of members, or for the single realization of the real world, yield highly uncertain results. Instead, we opt to neglect intermodel differences in the representation of internal variability and make use of the five LEs described in section 2 to estimate σ_x and $\sigma_{\Delta y}$ and, therefore, assume that the internal variability estimated from the five LEs is representative of that of the CMIP archive as a whole and of the observations. The validity of this will be discussed for each EC.

For a CMIP model with n_p and n_f ensemble members for the past and future, respectively, σ_x and $\sigma_{\Delta y}$ are estimated by subsampling n_p and n_f members (with replacement) from the past and future periods of each LE and repeating 1000 times. The LE ensemble mean is then subtracted from the mean of each subsample. The ensemble mean x and Δy for each LE will be much closer to the true $\overline{x}$ and $\Delta\overline{y}$ for that model than the subsamples when n_p and n_f are small, so these 1000 anomalies can be considered to represent 1000 deviations from the true $\overline{x}$ and $\Delta\overline{y}$ that could arise due to sampling of internal variability with only n_p and n_f members. When n_p or $n_f = 1$, these 1000 samples give no more information than would be obtained by using the individual members that make up the LE, but we follow the same procedure to allow all members from the LE to be used, while giving equal weighting to each LE.

The 1000 values for each LE are pooled together to give 5000 anomalies from the ''truth'' that have been sampled from five different models, each with their own representation of internal variability. The values of σ_x and $\sigma_{\Delta y}$ are then given by the standard deviation across these 5000 anomalies and $r_{x\Delta y}$ is simply the correlation between the LE subsamples used to calculate σ_x and $\sigma_{\Delta y}$. Thus, σ_x, $\sigma_{\Delta y}$, and $r_{x\Delta y}$ are assigned to each model depending on the model ensemble size and the variability from the five LEs. The σ_x for the observed value is calculated in the same way, using $n_p = 1$.

4. The Southern Hemisphere jet shift constraint (SHJET)

The SHJET constraint relates a model's past SH jet latitude (ϕ_o) to its future projected SH jet shift ($\Delta\phi$) under anthropogenic forcing. Kidston and Gerber (2010) first showed that, in the CMIP3 ensemble, in the annual mean, a model with a lower-latitude SH jet stream, exhibited a larger poleward shift by the end of the twenty-first century under anthropogenic forcing. A similar relationship was found by Son et al. (2010) for a model's poleward jet shift in response to ozone depletion. Simpson and Polvani (2016, hereafter SP2016) then revisited the conclusions of Kidston and Gerber (2010) in the CMIP5 archive and showed that the constraint was still present, but that it was actually marked by a strong seasonality, with the strongest correlation between ϕ_o and $\Delta\phi$ occurring during the SH winter. This seasonality also called into question the previously proposed hypothesis for why the SHJET constraint exists, namely that it is related to intermodel spread in eddy-feedback strength as identified through the southern annular model (SAM) time scale, since the intermodel spread in SAM time scale primarily occurs in the summer. So, this constraint still lacks explanation, but given that it has already been found in multiple model ensembles, we already have some confidence that it is robust to quasi out-of-sample testing. Indeed, Curtis et al. (2020) have already shown that the SHJET constraint is still present in CMIP6 over May–October. Here, we will draw the same conclusion but with a focus on the JJA season and further quantify the constraint on the future poleward shift of the SH jet that we may expect to see in the real world.

The jet latitude is defined as the latitude of the maximum 700-hPa zonal mean U in the SH, determined by finding the maximum of a quadratic fit using 700-hPa zonal mean zonal wind at three grid points on the 1° grid: the gridded maximum and the two adjacent grid points. The predictor for this constraint is the 1979–2014 JJA jet latitude (ϕ_o) and the predictand is the JJA future (2070–99) − past (1979–2014) jet shift ($\Delta\phi$), both expressed in degrees north.

Figure 2a reproduces the CMIP5 results of SP2016. A negative correlation between ϕ_o and $\Delta\phi$ exists: models with lower-latitude jets exhibit larger future poleward jet shifts. Furthermore, many of the models are biased equatorward relative to the reanalyses, suggesting that they may predict too large a poleward jet shift. Figure 2b now shows $\Delta\phi$ versus ϕ_o for the CMIP6 models. A negative correlation is still present in CMIP6 and it is still significant (Fig. 3c).

In CMIP6, the across-model variance in ϕ_o is reduced (Fig. 3a), although the variance in $\Delta\phi$ is similar (Fig. 3b). As would be expected under these circumstances, the fraction of variance explained by the EC has reduced in CMIP6 and it is the intermodel spread in other aspects of the forced response, not explained by the EC that is explaining relatively more of the variance. Depending on the method used, the EC explained 60.7%–62.4% of the variance in CMIP5 and explains 25.1%–25.6% of the variance in CMIP6[2] (Fig. 3d). For this EC, the choice of method affects the partitioning of variance explained between internal variability and the δ component (Fig. 3d) with less variance explained by internal variability in the BHM. This is due to the high correlation between errors in ϕ_o and $\Delta\phi$. There is a negligible contribution from intermodel

[2] Note that the variance explained here does not necessarily correspond to the square of the correlation coefficient shown in Fig. 3c because it reflects the variance explained across a single member from each model as opposed to across model ensemble means.

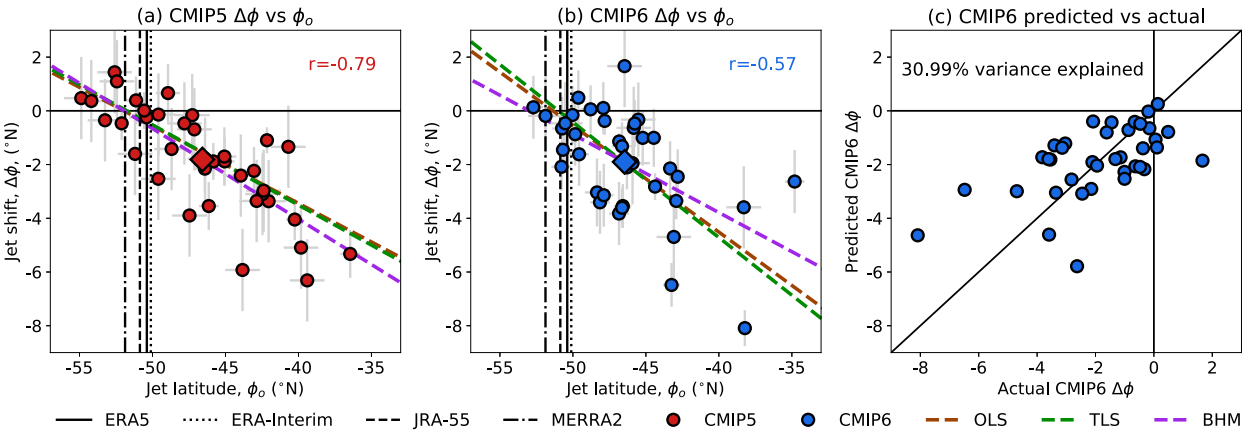

FIG. 2. (a) CMIP5 $\Delta\phi$ vs ϕ_o. Uncertainties for each model are derived using the method in section 3c and depicted here using the 95% confidence interval ($\pm 1.96\sigma$ range). Black vertical lines show ϕ_o for ERA5, ERA-Interim, JRA-55, and MERRA2. Brown, green, and purple dashed lines show the best fitting regression line for the OLS, TLS, and BHM methods, respectively. The correlation (r) is quoted in the top right. (b) As in (a), but for CMIP6. (c) The $\Delta\phi$ for CMIP6 that is predicted based on the CMIP5 BHM regression coefficients and the CMIP6 ϕ_o values vs the actual CMIP6 $\Delta\phi$ values. The percentage of CMIP6 variance in $\Delta\phi$ that is explained by the prediction is quoted.

spread in the globally averaged warming (Fig. 3d, black hatching on gray bars; see figure caption for method).

While the EC explains less variance in CMIP6, the relationship between ϕ_o and $\Delta\phi$ is still significant (Fig. 3c) and this still holds after crudely accounting for model interdependence by first averaging over models from the same modeling center (Fig. 1 in the online supplemental material). Furthermore, the α and β regression parameters are not particularly sensitive to the method used and agree between CMIP5 and CMIP6 within the uncertainties (Figs. 3e,f). This is further demonstrated in Fig. 2c where the CMIP5 BHM α and β are used to predict the CMIP6 $\Delta\phi$. This works reasonably well and indicates that we could have predicted 31.56% of the CMIP6 intermodel spread in ensemble mean $\Delta\phi$, based only on knowledge of the CMIP6 ϕ_o values and the CMIP5 EC. We can actually predict more of the CMIP6 variance with the CMIP5 regression coefficients than with those derived from CMIP6 itself, but this is likely just due to random chance.

Figures 4a and 4b indicate the extent to which CMIP6 can be considered an out-of-sample test compared to CMIP5 by

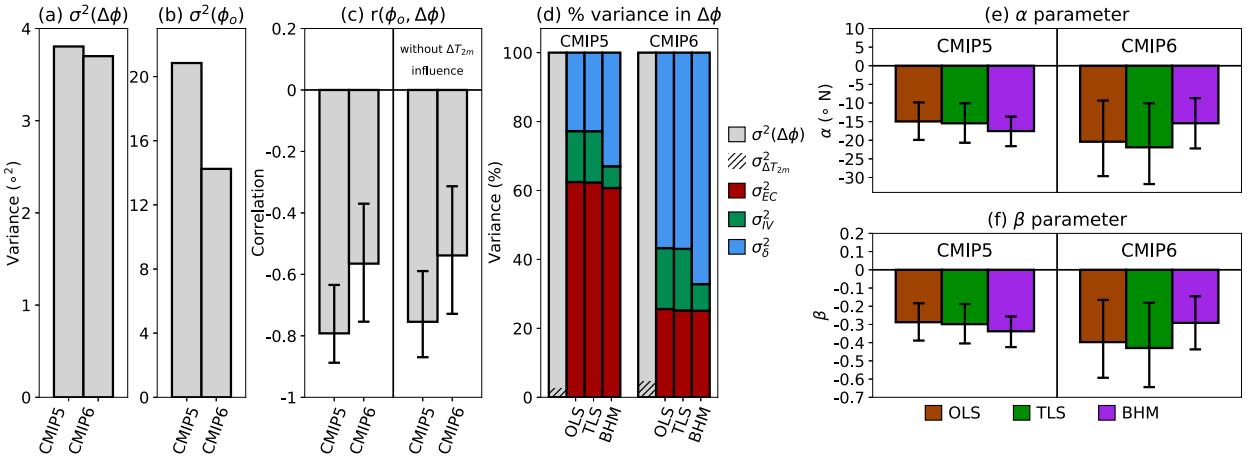

FIG. 3. (a) Variance in $\Delta\phi$ using ensemble means for each model. (b) As in (a), but for ϕ_o. (c) The correlation r between ϕ_o and $\Delta\phi$ both without (left) and with (right) first regressing out the component that is linearly related to globally averaged surface temperature change (ΔT_{2m}). Whiskers show the 95% confidence intervals estimated using a bootstrapping with replacement procedure. (d) A decomposition of the total variance across models, using a single member. The black hatching on the gray bar shows the percent variance explained by intermodel differences in ΔT_{2m}, calculated by differencing the total variance and the variance after regressing out the contribution that is linearly related to ΔT_{2m}. The colored bars show the percent variance explained by the EC (red), internal variability (green), and intermodel differences in the forced response that are unrelated to the EC (blue) for each method. (e),(f) The α and β regression parameters. Black ranges in (e) and (f) show the 95% confidence intervals derived by bootstrapping models with replacement for OLS and TLS and by using the 1000 estimates of the regression coefficients from the BHM.

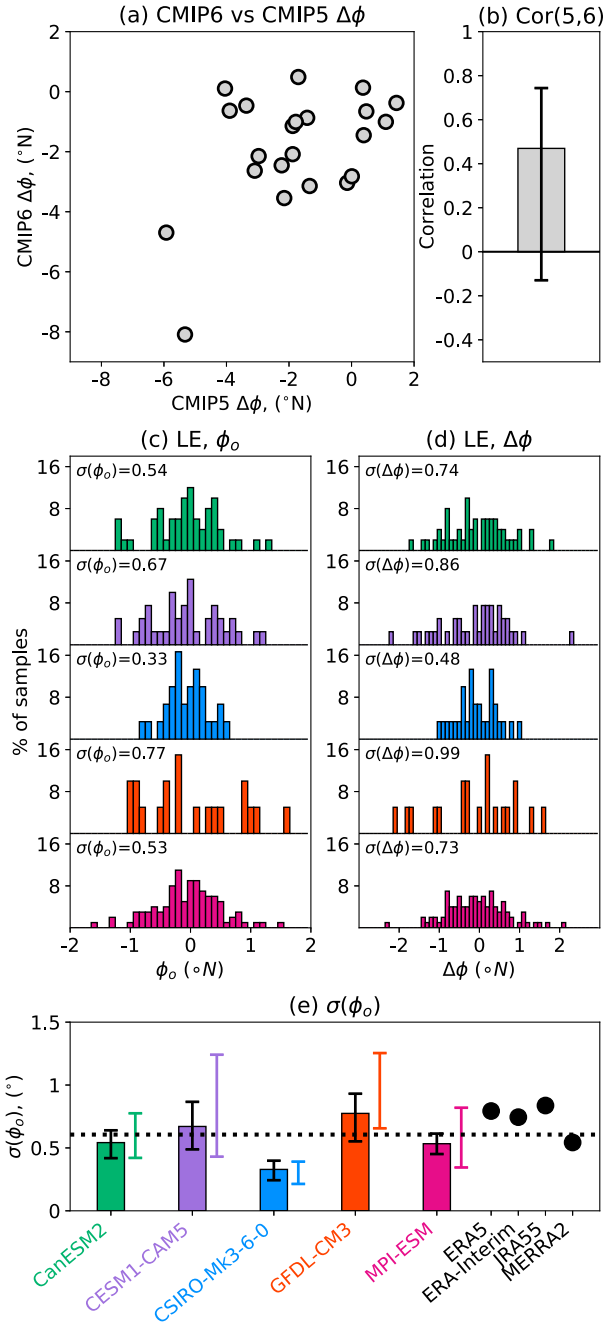

FIG. 4. (a) The relationship between $\Delta\phi$ in a CMIP6 model and its CMIP5 predecessor. (b) The correlation of the points shown in (a) along with a 95% confidence interval (estimated by bootstrapping models with replacement). (c) PDFs of the anomalies of ϕ_o from the true (ensemble mean) ϕ_o for each LE members for the five LEs, using the same color scheme as in (e). (d) As in (c), but for the distribution of $\Delta\phi$ anomalies from the ensemble mean. (e) Bars show the standard deviation of the distributions shown in (c). Black ranges show the uncertainty on the bars by bootstrapping with replacement the members in (c) and recalculating $\sigma(\phi_o)$. Colored ranges show values of $\sigma(\phi_o)$ calculated by bootstrapping the individual years from a single member, 1000 times, and calculating the standard deviation across these bootstrapped samples; $\sigma(\phi_o)$ is calculated this way for each member of the LE, and the range shows the range of values derived. Black points show an estimate of $\sigma(\phi_o)$ by bootstrapping years, with replacement for the reanalyses. The black dotted line shows the value of $\sigma(\phi_o)$ that is used to sample the observational uncertainty (i.e., that determined from the five LEs pooled together).

showing the correlation between $\Delta\phi$ in a CMIP6 model and $\Delta\phi$ in its CMIP5 predecessor using 22 models that are directly related (Table 1). The correlation between $\Delta\phi$ in the CMIP5 and CMIP6 models is not significant, although the correlation for ϕ_o in isolation is slightly higher and marginally significant (supplemental Fig. 2a). Overall, this suggests that we can consider CMIP6 to be at least a partial out-of-sample test compared to CMIP5.

To assess the extent to which the uncertainty on ϕ_o and $\Delta\phi$ (i.e., σ_x and $\sigma_{\Delta y}$ used in the regression) determined from the LEs might be representative for each CMIP model, the PDFs of ϕ_o and $\Delta\phi$ for the five LEs are shown in Figs. 4c and 4d along with a quantification of the standard deviation of ϕ_o across the members and its uncertainty in Fig. 4e (colored bars and black range). The closer the agreement between the LEs, the more confidence we may have that the values derived from them are representative for other CMIP models. There is a range in σ_{ϕ_o} but the black uncertainty ranges in Fig. 4e indicate that even with a large ensemble, it is still difficult to accurately determine σ_{ϕ_o}.

Another method that could have been considered for estimating σ_{ϕ_o} (and similarly $\sigma_{\Delta\phi}$) based on the data from a given model, as opposed to the LEs, is to bootstrap, with replacement, individual years from the past for a given model to generate a new climatology and a new estimate of ϕ_o. This could be repeated, say, 1000 times and σ_{ϕ_o} calculated as the standard deviation of ϕ_o across these 1000 samples. The colored ranges in Fig. 4e indicate the range of estimates of σ_{ϕ_o} determined this way from each individual LE member. This method has the potential to be highly inaccurate (cf. colored ranges with colored bars in Fig. 4e), presumably because 36 years within one member is not sufficient to truly characterize the full distribution of variability. So even though the value of σ_{ϕ_o} estimated from the LEs may not be truly representative for every model, it is likely more representative than what could be estimated from individual CMIP models when the ensemble size is small. Another method that could be considered for cases where there is no dependence of the internal variability on the climate state is to bootstrap from the preindustrial control simulations provided for each model, but we have not considered this here given that jet stream variability is expected to depend on climate state (Barnes and Polvani 2013).

To constrain future projections, we assume that σ_{ϕ_o} estimated from the LEs is representative of the uncertainty on the real world ϕ_o. The estimate of σ_{ϕ_o} from the reanalysis products using the bootstrapping of individual years method (black dots in Fig. 4e) is close to the estimate from the LEs (dotted line in Fig. 4e), especially considering the large

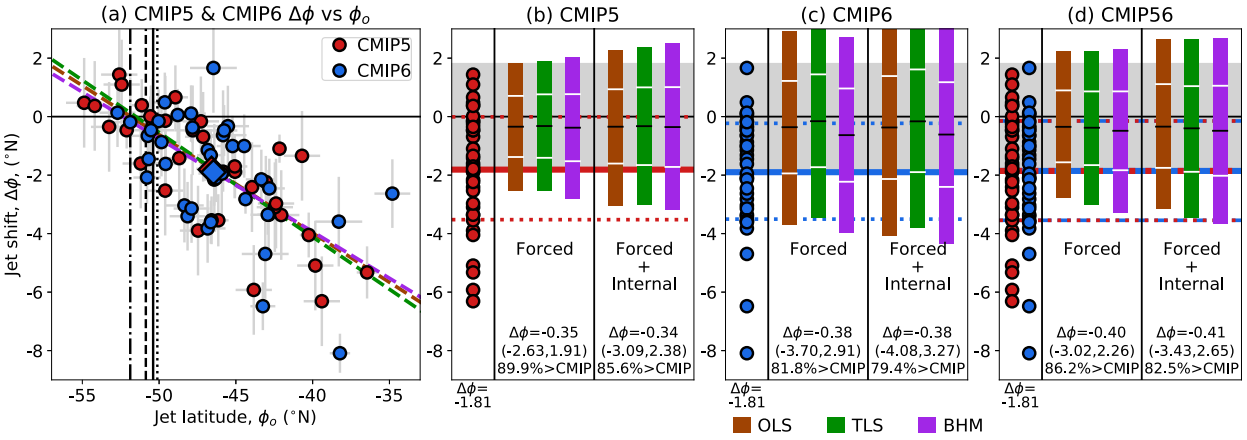

FIG. 5. (a) Jet shift ($\Delta\phi$) vs climatological jet latitude ϕ_o for ensemble means of the CMIP5 (red) and CMIP6 (blue) models along with best fitting regression lines for CMIP5 and CMIP6 combined for OLS (brown), TLS (green), and BHM (purple). Black vertical lines show the reanalyses (see the legend of Fig. 2) and filled diamonds show the CMIP5 and CMIP6 ensemble means. (b) CMIP5 constrained projections: the left portion reproduces the CMIP5 $\Delta\phi$ values of (a) along with the 66% confidence interval (dotted red lines); the middle portion shows the constrained projections for the forced response i.e., excluding internal variability; and the right portion shows the constrained projections for the forced response plus internal variability for a single realization. The red horizontal line across the panel shows the CMIP5 ensemble mean while the gray range shows the 95% confidence interval for the range of jet shifts that could arise due to internal variability for a single member, estimated from the LEs. (c) As in (b), but for CMIP6; (d) as in (b), but for CMIP5 and CMIP6 combined. Brown, green, and purple ranges with black line and white range show the 95% confidence intervals, mean, and 66% confidence intervals of constrained projections using TLS, OLS, and BHM, respectively. Values quoted are from the mean across the methods and are, from top to bottom, $\Delta\phi$, the range of the 95% confidence interval, and the probability of the real world future $\Delta\phi$ being less poleward than the relevant CMIP ensemble mean.

uncertainty associated with this method, suggesting this is a reasonable approach.

Constraining projections for $\Delta\phi$ following the methods in section 3 indicates that the poleward shift in the real world will be considerably smaller than the CMIP5 or CMIP6 multimodel mean, since the reanalysis jet position is at the poleward end of the CMIP model distribution (Fig. 5). Note that, unlike Curtis et al. (2020), we do not find a large reduction in the CMIP6 multimodel mean poleward shift compared to CMIP5, but there are a variety of differences in our methods, including a focus on a different forcing scenario (they used 4xCO2 simulation) and a narrower winter season (they used May–October). Also, unlike Bracegirdle et al. (2020) we do not find a substantial improvement in ϕ_o in CMIP6. While our methods differ, we suspect this may be largely due to the models considered as the two most equatorward models in CMIP6 here (MIROC-ES2L and CNRM-CM6-1-HR) were not included in that study. Taking the mean across the CMIP5 and CMIP6 models gives a projected jet shift of 1.81° poleward while that for the constrained jet shift distribution using OLS, TLS, and BHM is 0.35°, 0.38°, and 0.48° poleward, respectively, for the forced response, with 95% confidence intervals of (2.8° poleward to 2.2° equatorward), (3.0° poleward to 2.2° equatorward), (3.3° poleward to 2.3° equatorward) with an 88.5%, 86.5%, 83.4% chance, respectively, that the forced poleward shift in the real world will be less than the CMIP5 and CMIP6 ensemble mean. The three regression methods give similar results, but slightly larger differences are seen for CMIP6 than CMIP5, perhaps because the constraint explains less variance in CMIP6, leaving more room for the regression methods to differ.

When also considering the role that internal variability may play in our one potential future (right portion of Figs. 5b–d), the 95% confidence intervals suggest that it is very unlikely we will observe a poleward shift of more than around 3.5°. In fact, following Cox et al. (2018), if we consider the 66% confidence interval (white lines on each colored bar) to correspond to the "likely range" according to Intergovernmental Panel on Climate Change (IPCC) definitions, we find the CMIP ensemble mean sits at the very poleward edge of this likely range.

Overall, the SHJET emergent constraint still survives in CMIP6 and it could represent a useful constraint on the poleward shift we should expect to see in the real world, if the mechanism behind the constraint could be fully understood. Toward that end, supplemental Fig. 3 repeats another component of the SP2016 analysis and indicates that fully understanding this constraint could likely be achieved by understanding why, in winter, the forced changes in zonal wind are roughly anchored to the same position in each model, regardless of climatological jet latitude differences, such that the wind anomalies lead to a poleward shift of low-latitude jets but a strengthening for higher-latitude jets.

5. North American stationary waves with extension to regional hydroclimate (VWIND)

Our second EC aims at constraining meridional wind (V) changes over North America during the December–February (DJF) season and was proposed by Simpson et al. (2016, hereafter S2016). The CMIP5 ensemble mean change in 300-hPa V consisted of southerlies off the U.S. west coast, northerlies

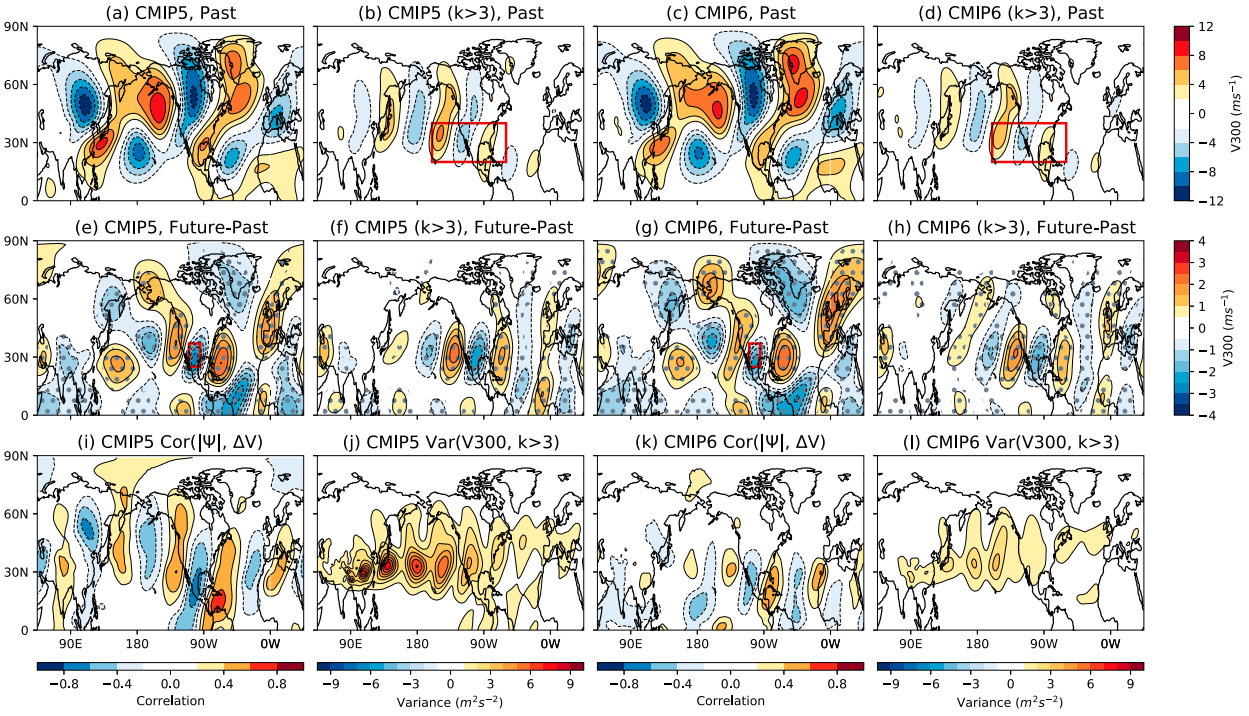

FIG. 6. A comparison of 300-hPa meridional wind between (left) CMIP5 and (right) CMIP6. (a),(c) The climatological V300 for CMIP5 and CMIP6, respectively; (b),(d) as in (a) and (c), but after filtering to only retain zonal wavenumbers greater than 3. (e),(g) The future − past difference for CMIP5 and CMIP6, respectively; (f),(h) as in (e) and (g), but after filtering to only retain zonal wavenumbers greater than 3. (i),(k) The correlation between the climatological $k > 3$ stationary wave amplitude ($|\psi|$) and the change in 300-hPa meridional wind across models for CMIP5 and CMIP6. (j),(l) The across-model variance in $k > 3$ meridional wind for CMIP5 and CMIP6, respectively. The stippling in (e)–(h) shows regions where more than 80% of the models agree on the sign of the anomaly. The red box in (b) and (d) shows the region used to define $|\psi|$ (160°–60°W, 20°–40°N). The red box in (e) and (g) shows the region used to define ΔV_{sw} (110°–95°W, 25°–37°N).

over the U.S. interior southwest, and southerlies off the U.S. east coast (Fig. 6e). The same pattern is also found in CMIP6 (Fig. 6g). These V anomalies are primarily associated with intermediate scale (zonal wavenumber $k > 3$) stationary waves that are meridionally trapped in the Pacific waveguide (Figs. 6b and 6d), as can be seen by comparing the full change in V with that after filtering for $k > 3$ (compare Figs. 6e and 6f, and Figs. 6g and 6h). By reproducing the V changes in a stationary wave model when only imposing changes in the upper tropospheric zonal mean zonal wind, and using stationary wave theory, S2016 argued that the mechanism behind this change involves the warming-induced strengthening of the westerlies in the subtropical upper troposphere, acting to lengthen the scale of stationary waves that can be supported by the Pacific waveguide, leading to the V anomalies downstream over North America. Based on this mechanistic understanding, it would then be expected that the magnitude of a model's change in V would be related to (i) the amplitude of a model's intermediate scale stationary waves in the Pacific–North American sector and (ii) the strengthening of the zonal mean westerlies in the subtropical upper troposphere (these two quantities are uncorrelated).

S2016, therefore, used two predictors to predict the change in eddy meridional wind averaged over the interior southwest of the United States (ΔV_{sw}; red box in Fig. 6e): (i) the root-mean-square amplitude of the $k > 3$ stationary waves in the past ($|\psi|$), based on eddy V over a region in the eastern Pacific/southern United States (red box in Fig. 6b); and (ii) the change in zonal mean zonal wind at 100 hPa averaged from 20° to 40°N (ΔU_{100}). Consistent with the proposed mechanism, ΔV_{sw} was found to be significantly negatively correlated with $|\psi|$; that is, a model with larger amplitude $k > 3$ stationary waves in its climatology exhibited larger northerly anomalies over the interior southwest United States (Figs. 7a and 8c) and a larger amplitude of the meridional wind pattern over North America more generally (Fig. 6i). A negative, although insignificant, correlation was also found between ΔV_{sw} and ΔU_{100}: the larger the increase in zonal mean zonal wind in the subtropical upper troposphere, the larger the negative anomaly in ΔV_{sw} (Fig. 7b). A multiple linear regression using ΔU_{100} and $|\psi|$ as predictors explained a substantial fraction of the variance in ΔV_{sw} (Fig. 7c). S2016 argued that this was further evidence for their proposed mechanism; in addition, since many of the models have stationary wave amplitudes ($|\psi|$) that are too large (compare with the reanalyses in Fig. 7a), they inferred that the real world ΔV_{sw} will likely be smaller than the CMIP5 ensemble mean.

While the VWIND constraint was accompanied by a plausible mechanism that was verified by stationary wave model

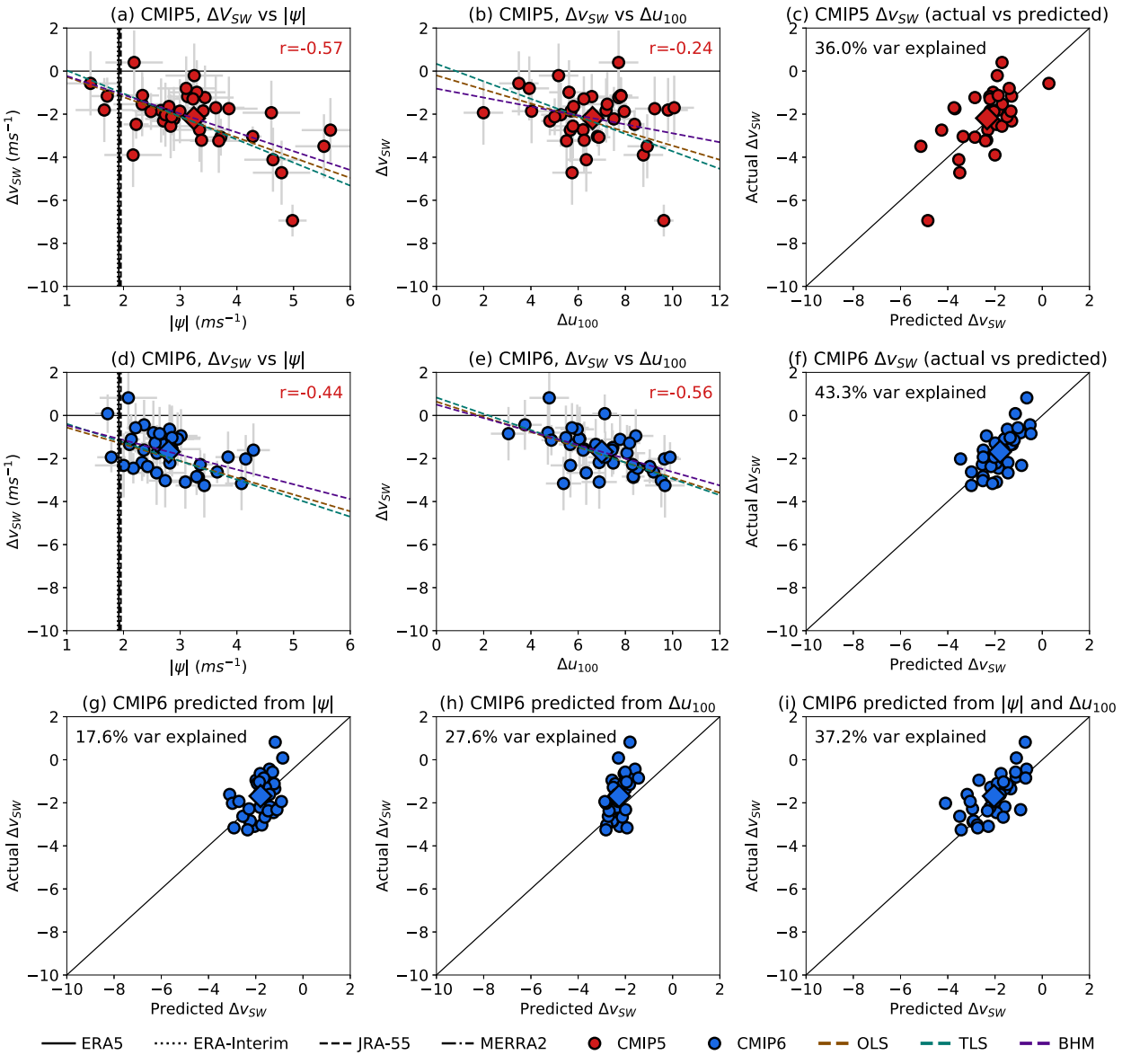

FIG. 7. (a) ΔV_{sw} vs $|\psi|$, (b) ΔU_{100} vs $|\psi|$, and (c) the relationship between the actual ΔV_{sw} and that predicted using multiple linear regression onto $|\psi|$ and ΔU_{100}, all for CMIP5. Uncertainties for each model in (a) and (b) are derived using the method in section 3c and depicted here using the 95% confidence interval ($\pm 1.96\sigma$ range). Black vertical lines in (a) show $|\psi|$ for ERA5, ERA-Interim, JRA-55, and MERRA2. Brown, green, and purple dashed lines show the best fitting regression line for OLS, TLS, and BHM. (d)–(f) As in (a)–(c), but for the CMIP6 models. (g) The CMIP6 Δv_{sw} that is predicted based on the CMIP5 BHM regression onto $|\psi|$ vs the actual CMIP6 Δv_{sw} values. (h) The Δv_{sw} for CMIP6 that is predicted based on the CMIP5 BHM regression onto Δu_{100} vs the actual Δv_{sw}. (i) The CMIP6 Δv_{sw} that is predicted based on the CMIP5 multiple linear (OLS) regression onto $|\psi|$ and Δu_{100} vs the actual Δv_{sw}. Filled diamonds in each panel show the ensemble mean.

experiments and also was shown to be robust in perturbed physics experiments by van Niekerk et al. (2017), we now test whether it still survives in CMIP6. It is only the predictor $|\psi|$ that can be used for an emergent constraint since ΔU_{100} relies on future information. We do, however, show the relationship between ΔV_{sw} and ΔU_{100} in Fig. 7 as well, to lend support to the stationary wave theory argument of S2016, given that it is important to accompany an EC with a mechanistic explanation. Negative correlations between ΔV_{sw} and both $|\psi|$ and ΔU_{100} are still found in CMIP6 (Figs. 7d,e). The correlation between $|\psi|$ and ΔV_{sw} is significant for both CMIP5 and CMIP6 (Fig. 8c) and this remains true after crudely accounting for model interdependence by first averaging over models from the same modeling center (supplemental Fig. 4). Lending support to the stationary wave theory arguments of S2016, combining $|\psi|$ and ΔU_{100} in a multiple linear regression explains a similar fraction of variance in CMIP6 as it does in CMIP5 regardless of which ensemble the regression coefficients are derived from

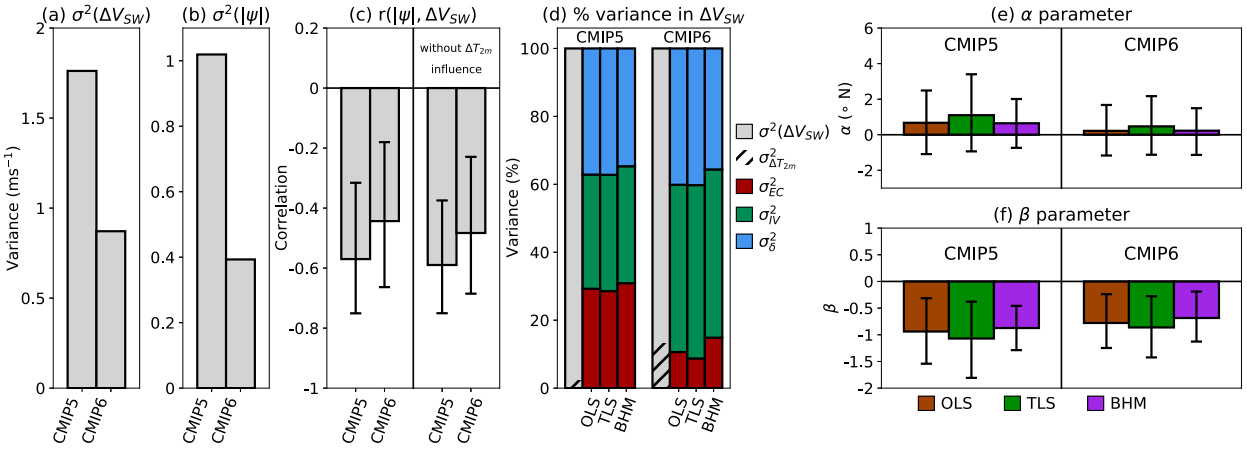

FIG. 8. (a) Variance in ΔV_{sw} using the ensemble mean for each model. (b) As in (a), but for $|\psi|$. (c) The correlation between $|\psi|$ and ΔV_{sw} both without (left) and with (right) first regressing out the component that is linearly related to globally average surface temperature change ΔT_{2m}. Whiskers show the 95% confidence interval derived by bootstrapping models with replacement. (d) A decomposition of the total variance in ΔV_{sw} across models using a single member. The black hatching on the gray bar shows the percent variance explained by intermodel differences in ΔT_{2m}, calculated by differencing the total variance and the variance after regressing out the contribution that is linearly related to ΔT_{2m}. The colored bars show the percent variance explained by the EC (red), internal variability (green), and intermodel differences in the forced response that are unrelated to the EC (blue) for each method. (e),(f) The α and β regression parameters for $\Delta V_{sw} = \alpha + \beta|\psi|$. Black ranges in (e) and (f) show the 95% confidence interval derived by bootstrapping models with replacement for OLS and TLS and by using the 1000 regression coefficients estimated from the BHM.

(Figs. 7f and 7i). Focusing now on the only aspect that can be used as an emergent constraint—the relationship between ΔV_{sw} and $|\psi|$ according to $\Delta V_{sw}(i) = \alpha + \beta|\psi|(i) + \varepsilon(i)$—using the CMIP5 regression coefficient based on $|\psi|$ alone to predict the CMIP6 ΔV_{sw} can only explain 16% of the variance (Fig. 7g).

The variance across models in ΔV_{sw} is reduced by 0.87 $m^2 s^{-2}$ in CMIP6 compared to CMIP5 (Fig. 8a). This is expected based on the EC given that the variance across models in $|\psi|$ ($\sigma^2_{|\psi|}$) is also substantially reduced by roughly 0.63 $m^2 s^{-2}$ (Fig. 8b; cf. Figs. 6l and 6j). In CMIP6 (Fig. 7d), there are no longer models on the extreme biased end of the CMIP5 range in $|\psi|$ (cf. Figs. 7a and 7d), resulting in only 10.6%, 8.7%, and 14.9% of the variance is ΔV_{sw} being explained by the EC, compared to 29.2%, 28.6%, and 30.9% in CMIP5 for OLS, TLS, and BHM, respectively (Fig. 8d). So, even though the regression coefficients that relate ΔV_{sw} to $|\psi|$ are similar between CMIP5 and CMIP6 (Figs. 8e,f), because of the reduced spread in the predictor the EC is less effective in CMIP6, as will be further demonstrated below.[3]

Before quantifying the constraint on ΔV_{sw}, we first check some of our assumptions. First, ΔV_{sw} in the CMIP6 models is not correlated with ΔV_{sw} in its CMIP5 predecessor (Figs. 9a and 9b). The stationary wave amplitude $|\psi|$ in the CMIP6 models is correlated with that in CMIP5 but they do not follow the 1:1 line (supplemental Fig. 5). Overall, this suggests it is reasonable to consider CMIP6 as being at least a partial out-of-sample test compared to CMIP5. Second, there are only minor differences across the LEs in the uncertainty in $|\psi|$ and ΔV_{sw} (Figs. 9c–e) suggesting that, while assigning a $\sigma_{|\psi|}$ and $\sigma_{\Delta V_{sw}}$ to each CMIP model based on the LEs may not be completely accurate, it is likely a best estimate in the absence of a large ensemble for each CMIP model. Finally, the estimated $\sigma_{|\psi|}$ from the observations obtained by bootstrapping individual years, while likely subject to considerable uncertainty (colored ranges in Fig. 9e), is close to the estimate from the LEs (cf. black dots and dotted line in Fig. 9e), indicating that basing $\sigma_{|\psi|}$ of the real world on the LEs is likely a reasonable approximation.

The constraint on ΔV_{sw} is quantified in Fig. 10 using the methods of section 3. In CMIP5, the EC does indeed represent a substantial constraint on ΔV_{sw} with OLS, TLS, and BHM indicating an 84.4%, 84.9%, and 86.8% chance, respectively, that the forced response will be smaller (less northerly) than the CMIP5 ensemble mean along with a constrained mean ΔV_{sw} of −1.17, −1.16, and −1.04 m s^{-1}, respectively, compared to −2.19 m s^{-1} in the CMIP5 ensemble mean (Fig. 10b). The constrained distribution of forced ΔV_{sw} here is broader than that in S2016 since they did not incorporate uncertainty in the observed value of $|\psi|$, or other contributions to the forced response, except for the influence of ΔU_{100}. The constrained distribution of ΔV_{sw} from CMIP5 that includes the influence of internal variability is broader still, but nevertheless still indicates that the ΔV_{sw} we could expect to see in the real world has around an 83% chance of being smaller than the CMIP5 ensemble mean (Fig. 10b).

[3] The models that had large biases in CMIP5 were BCC-CSM-1-m with $|\psi|$ = 5.7, which is much improved in its CMIP6 successor BCC-CSM2-MR ($|\psi|$ = 3.1); MRI-CGCM3 ($|\psi|$ = 4.6), which has improved to $|\psi|$ = 3.8 in MRI-ESM2-0; FGOALS-g2 ($|\psi|$ = 4.6), which has improved slightly to $|\psi|$ = 4.3 in FGOALS-g3; and the IPSL variants, which ranged from $|\psi|$ = 4.8 to 5.5 and have now improved to $|\psi|$ = 4.1 in IPSL-CM6A-LR.

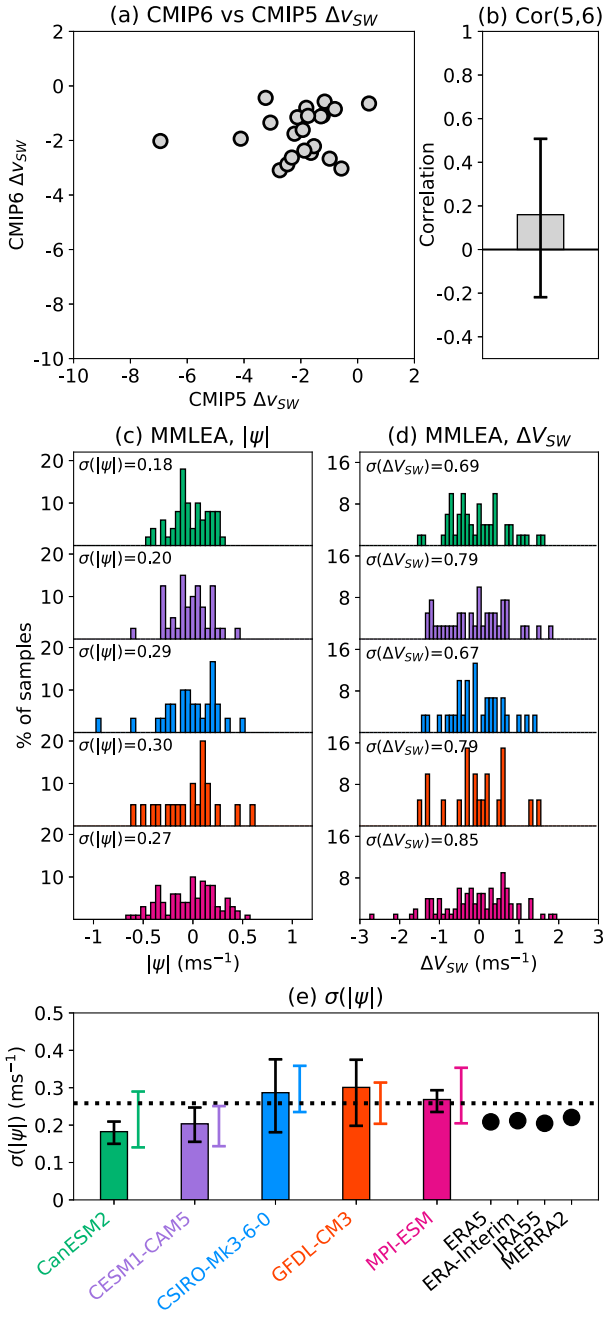

FIG. 9. (a) The relationship between Δv_{sw} in a CMIP6 model and its CMIP5 predecessor. (b) The correlation of the points shown in (a) along with 95% confidence interval derived by bootstrapping models with replacement. (c) PDFs of the anomalies in stationary wave amplitude $|\psi|$ from the true (ensemble mean) $|\psi|$ for each LE member for the five LEs, using the same color scheme as in (e). (d) As in (c), but for the distribution of Δv_{sw} for future − past differences from each LE members. (e) Bars show the standard deviation of the distributions shown in (c). Black ranges show the uncertainty on the bars by bootstrapping with replacement the members of the LE and recalculating $\sigma|\psi|$. Colored ranges show values of $\sigma|\psi|$ calculated by bootstrapping the individual years from a single member, 1000 times, and calculating the standard deviation across these bootstrapped samples. The $\sigma|\psi|$ is calculated this way for each member of the LE and the range shows the range of values obtained. Black points show an estimate of $\sigma|\psi|$ for the reanalysis using the bootstrapping method. The black dotted line shows the value of $\sigma_{|\psi|}$ that is used to sample the observational uncertainty i.e., that determined from the five LEs pooled together.

Figure 10a indicates that the ensemble mean shift to a smaller ΔV_{sw} in CMIP6 compared to CMIP5 is consistent with a shift along the regression line accompanying a smaller ensemble mean $|\psi|$. However, the fact that CMIP6 no longer contains models with extremely large negative values of ΔV_{sw} means that the EC is no longer really an effective constraint (Fig. 10c). The constrained range, including internal variability, encompasses almost all the CMIP6 models, although it does still suggest the most likely value of ΔV_{sw} is slightly smaller than the ensemble mean.

A constraint on upper-level V may be of limited practical use by itself, but S2016 demonstrated that these V changes have an equivalent barotropic structure and the accompanying near surface meridional wind anomalies have implications for regional hydroclimate. They did not quantify the associated constraint on precipitation, but argued that we should expect the real world to behave like models with smaller $|\psi|$, which exhibited less wetting over the U.S. west coast and less drying over the interior southwest than the CMIP5 ensemble mean.

Here, this analysis is extended to provide a more rigorous quantification of the precipitation constraint (or lack thereof) in three regions: the U.S. west coast, the U.S. south, and southern Mexico (Fig. 11a, red). Note that ΔP in these regions is clearly correlated with $|\psi|$ across the CMIP5 models in the manner described by S2016 (Fig. 11a). Models with large-amplitude stationary waves exhibit more wetting on the U.S. west coast and more drying over the interior southwest and Mexico associated with their larger meridional wind changes. In CMIP6, however, these correlations between $|\psi|$ and precipitation are largely absent, except over southern Mexico (Figs. 11b,e,h,k).

The reason for the disappearance of this correlation structure in CMIP6 is because there are no longer any models that have very large $|\psi|$ and it was those models that were dominating in the CMIP5 correlations. This becomes clear by comparing the CMIP5 correlations with those after omitting the CMIP5 models that have $|\psi|$ larger than the maximum $|\psi|$ in CMIP6 (compare red solid and hatched bars in Figs. 11e,h,k). Once these six CMIP5 models have been omitted, the correlation between $|\psi|$ and precipitation in these regions is no longer significant in CMIP5 either.

In CMIP5, even though there was a significant correlation between $|\psi|$ and west coast precipitation, it was never really that effective a constraint, as there were too many additional uncertainties (Fig. 12b) and the same is true now in CMIP6 (Fig. 12c). Over the southern United States, the CMIP5 constraint suggested the real world response will more likely be a slight wetting as opposed to the slight drying seen in the CMIP5

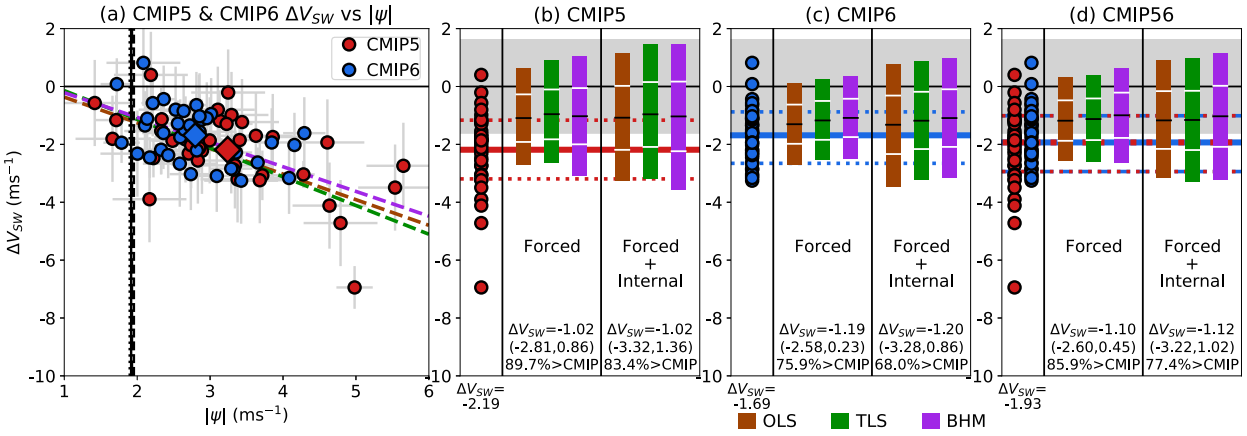

FIG. 10. (a) CMIP5 (red) and CMIP6 (blue) ΔV_{sw} vs $|\psi|$ along with the best fitting regression lines for CMIP5 and CMIP6 combined using OLS (brown), TLS (green), and BHM (purple). The four reanalyses are shown by the black lines (see Fig. 7 legend). (b) The left portion of the panel reproduces the CMIP5 ΔV_{sw} from (a) and the red line that spans the panel shows the CMIP5 ensemble mean. The gray shaded region shows the 95% confidence interval of anomalies in ΔV_{sw} that could arise due to internal variability, determined from the LEs. In the middle portion of the panel, the colored bars show the 95% confidence interval of the constrained forced response using OLS (brown), TLS (green), and BHM (purple). The black line shows the mean of the constrained distribution and white lines delineate the 66% confidence interval on the constrained distribution. The right portion of the panel is the same as the middle but including the contribution to ΔV_{sw} from internal variability that could arise in our one future. Numbers quoted are, from top to bottom, the mean across the methods of the mean change in ΔV_{sw}, the 95% confidence interval of the constrained distribution (ci), and the probability that Δv_{sw} will be greater than (less negative) than the CMIP ensemble mean. (c),(d) As in (b), but for CMIP6 and for CMIP5 and CMIP6 combined, respectively.

ensemble mean and that those models with a very extreme drying are very unlikely (Fig. 12f). Consistent with this, in CMIP6 now that the stationary wave biases have been reduced, the ensemble mean precipitation change has shifted to being slightly positive and there are no longer any models with an extreme drying (Fig. 12g). However, the reduced spread in CMIP6 does also mean that, with the uncertainties involved, the constraint does not narrow down projected changes beyond the CMIP6 distribution. Over southern Mexico, the CMIP5 EC indicated that we should expect to see less drying than the CMIP5 ensemble mean (Fig. 12j). Indeed, with the model improvements in CMIP6, there is slightly less drying over this region (Fig. 12k). The EC is still somewhat useful here and suggests that there is a reasonable chance (~72%) that the real world will not exhibit as much forced drying as the CMIP6 ensemble mean (Fig. 12k). However, the constrained range incorporating internal variability is still sufficiently wide that it almost encompasses all of the CMIP6 models.

6. The ENSO-based EC on projected California precipitation change (CALP)

Another emergent constraint on DJF California precipitation was proposed through CMIP5 by Allen and Luptowitz (2017), AL2017, hereafter. They related a model's 2006 to 2100 trend in DJF California precipitation to its representation of the interannual correlation between DJF ENSO variability (given by the Niño-3.4 index) and DJF California precipitation. The proposed mechanism behind this constraint was that the future projected California precipitation changes were related to the El Niño–like SST warming trend seen in many models and that models with more realistic interannual ENSO teleconnections are more likely to simulate this forced change correctly.

AL2017 found that the real world correlation between ENSO and California precipitation using winters from 1948/49 to 2014/15 was 0.36 and argued that many models do not accurately represent this, with model values ranging from −0.12 to 0.58. They generated two groups of models based on their correlation between Niño-3.4 and California precipitation, r(Niño, pr). The models with a correlation below 0.2 were referred to as the "LOW-r" models and those with a correlation above 0.3 were referred to as the "HIGH-r" models, with the HIGH-r models considered more realistic. It was found that the average California precipitation trends were higher in the HIGH-r models. While AL2017 did not explicitly quantify the constraint on California precipitation, they argued that the models that exhibit more realistic ENSO teleconnections to California, exhibited larger and more consistent increases in California precipitation over the twenty-first century. One was left to infer that we should expect the real world to behave more like these models. This is the opposite conclusion to the constraint drawn by S2016 above, which was obviously problematic at the time. However, it has been shown above that, while the mechanism of S2016 shows promise, it does not represent an effective constraint on U.S. west coast precipitation, given the uncertainties. So, we see now whether the AL2016 constraint does any better.

Figure 13a shows r(Niño-3.4, pr). Six estimates of the observed correlation over the 1948 to 2014 period are shown using all combinations of datasets described in section 2 and these range from 0.30 to 0.35. However, there is likely a large

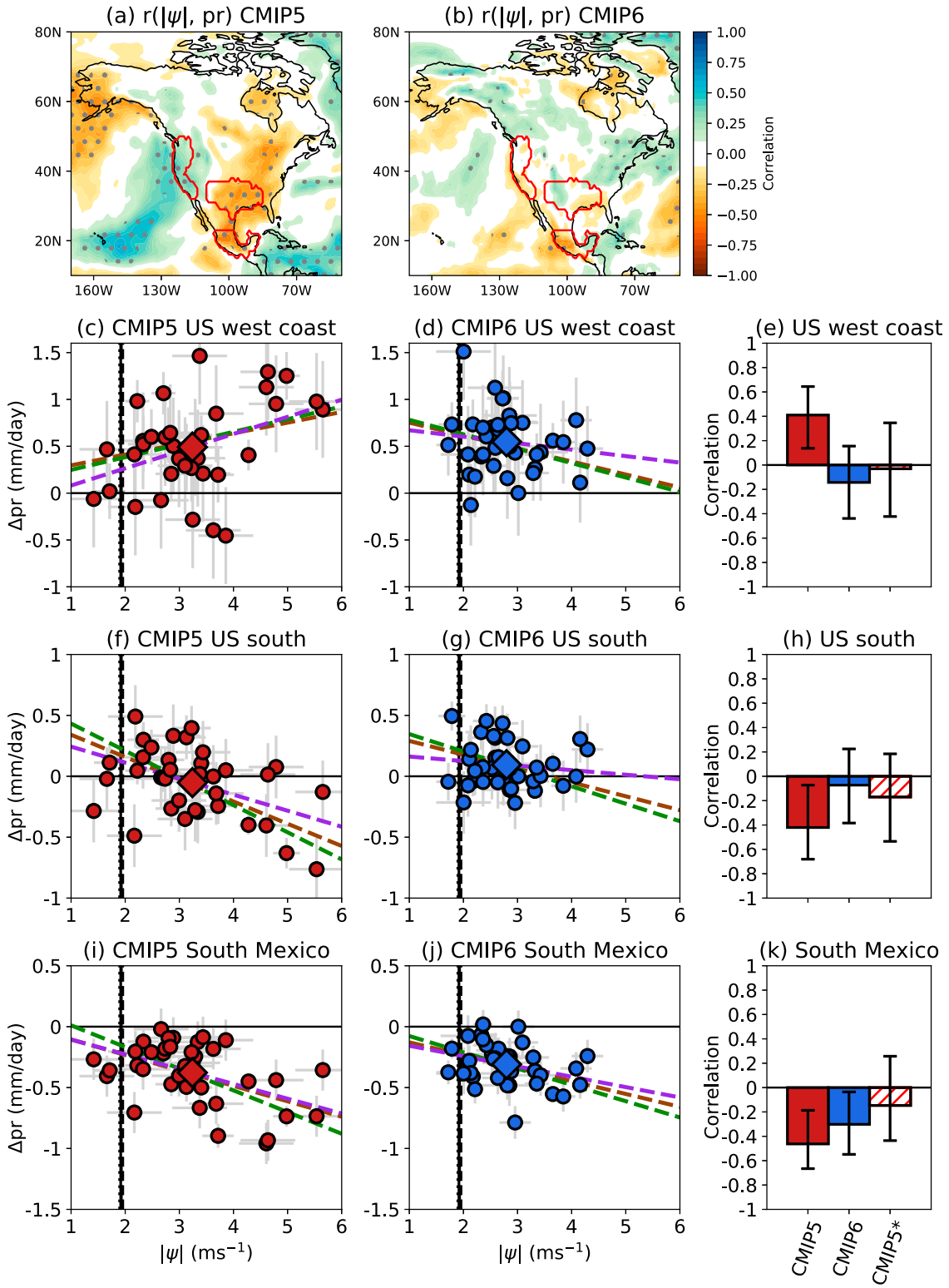

FIG. 11. (a),(b) The correlation between the stationary wave amplitude $|\psi|$ and Δpr for CMIP5 and CMIP6 respectively. Stippled regions are significant at the 95% confidence level by a two-sided bootstrapping test and red contours depict the regions used in (c)–(k). (c),(d) The relationship between Δpr averaged over the U.S. west coast and $|\psi|$ for CMIP5 and CMIP6 respectively along with $|\psi|$ for the reanalyses and the best fitting linear regression (brown = OLS, green = TLS, purple = BHM). (e) The red and blue bars display the correlation and its 95% confidence interval (derived by bootstrapping models with replacement) of the points shown in (c) and (d), respectively, and the red hatched correlation, referred to as CMIP5*, shows the correlation across the CMIP5 models, excluding those that have $|\psi|$ larger than the maximum $|\psi|$ in CMIP6. (f)–(h) As in (c)–(e), but averaged over the U.S. South; (i)–(k) as in (c)–(e), but averaged over southern Mexico.

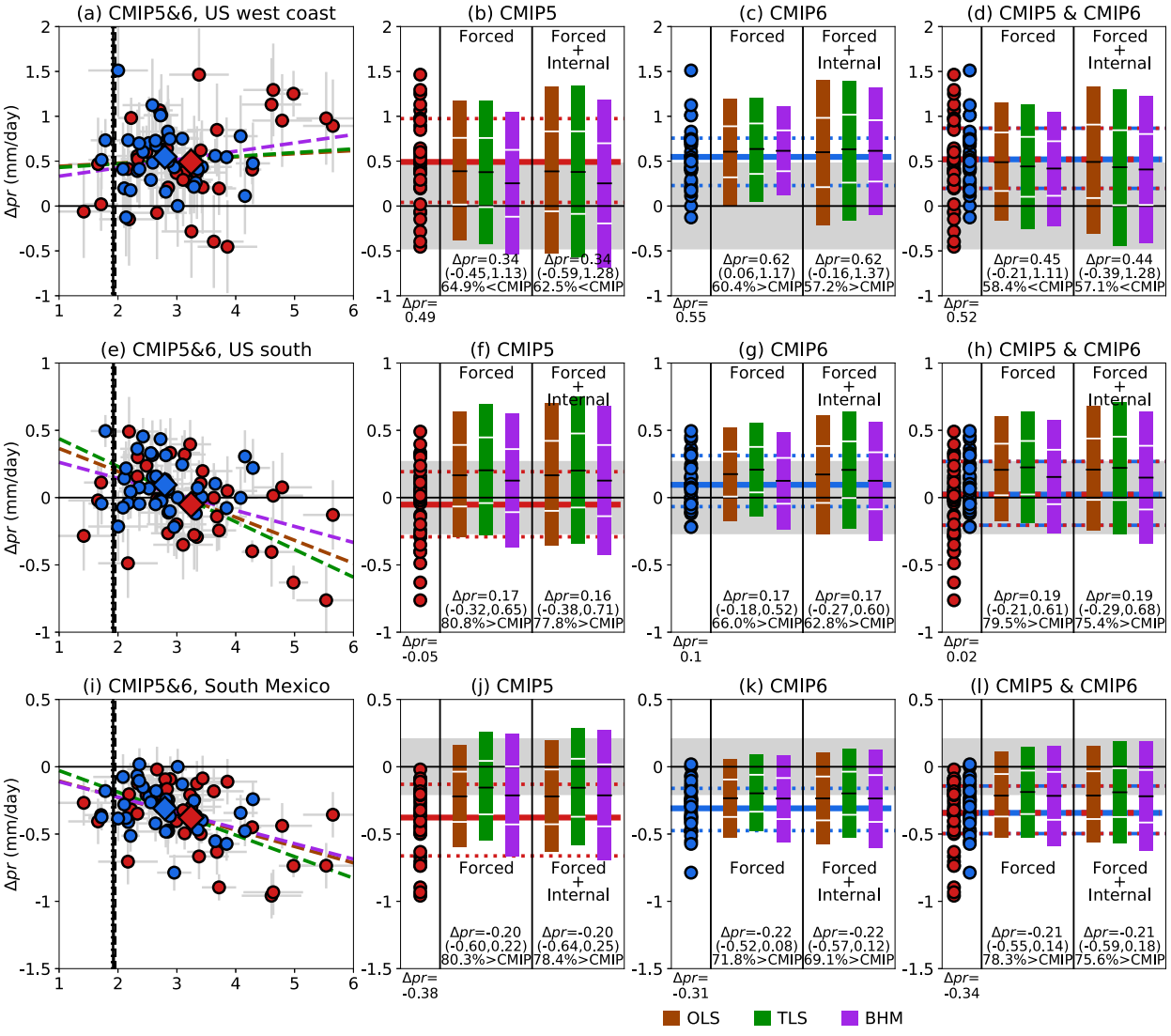

FIG. 12. (a) The relationship between Δpr averaged over the U.S. west coast and $|\psi|$ (red circles = CMIP5, blue circles = CMIP6, red diamond=CMIP5 mean, blue diamond = CMIP6 mean, black vertical lines = the reanalyses $|\psi|$, brown green, and purple dashed lines = the OLS, TLS, and BHM regression lines using CMIP5 and CMIP6 combined). (b) The CMIP5 Δpr averaged over the U.S. west coast (shown at left), the constrained distributions of the forced change (shown at center), and the constrained distribution of the forced change + internal variability (shown at right). Colored bars show the 95% confidence interval of the constrained distribution; the black line shows the mean, and white range shows the 66% confidence interval. Red lines spanning the panel show the CMIP5 ensemble mean and 66% confidence interval, and the gray range shows the 95% confidence range of Δpr values that could arise due to internal variability. (c) As in (b), but for CMIP6; (d) as in (b), but for CMIP5 and CMIP6 combined. (e)–(h) As in (a)–(d), but averaged over the U.S. south. (i)–(l) As in (a)–(d), but averaged over southern Mexico. The values quoted are, from top to bottom, the mean across the methods of the mean Δpr, the 95% confidence interval of the constrained distribution (ci), and the probability, from the constrained distribution, that the Δpr will be greater than or less than the CMIP5 or CMIP6 ensemble mean (whichever is larger).

uncertainty in this value due to internal variability. The gray shaded range shows the 95% confidence interval on r(Niño-3.4, pr) using GPCC and ERSSTv5, estimated by bootstrapping the individual DJF seasons from 1948–2014 with replacement and recalculating r(Niño-3.4, pr) 1000 times. This ranges from 0.04 to 0.51 and this magnitude of uncertainty is supported by the range of values estimated from 1948 to 2014 of the LEs (right side of Fig. 13a). As such, there is little motivation for choosing the threshold of r(Niño-3.4, pr) < 0.2 to identify models that do not represent r(Niño-3.4, pr) well.

Nevertheless, we proceed to reproduce the results of AL2017 by comparing 2006–99 California precipitation trends between models with r(Niño-3.4, pr) > 0.3 (HIGH-r) and models with r(Niño-3.4, pr) < 0.2 (LOW-r). First, using r(Niño-3.4, pr) calculated over 2006–99 after detrending each field, as in AL2017, we are left with 15 models in each group, with substantial

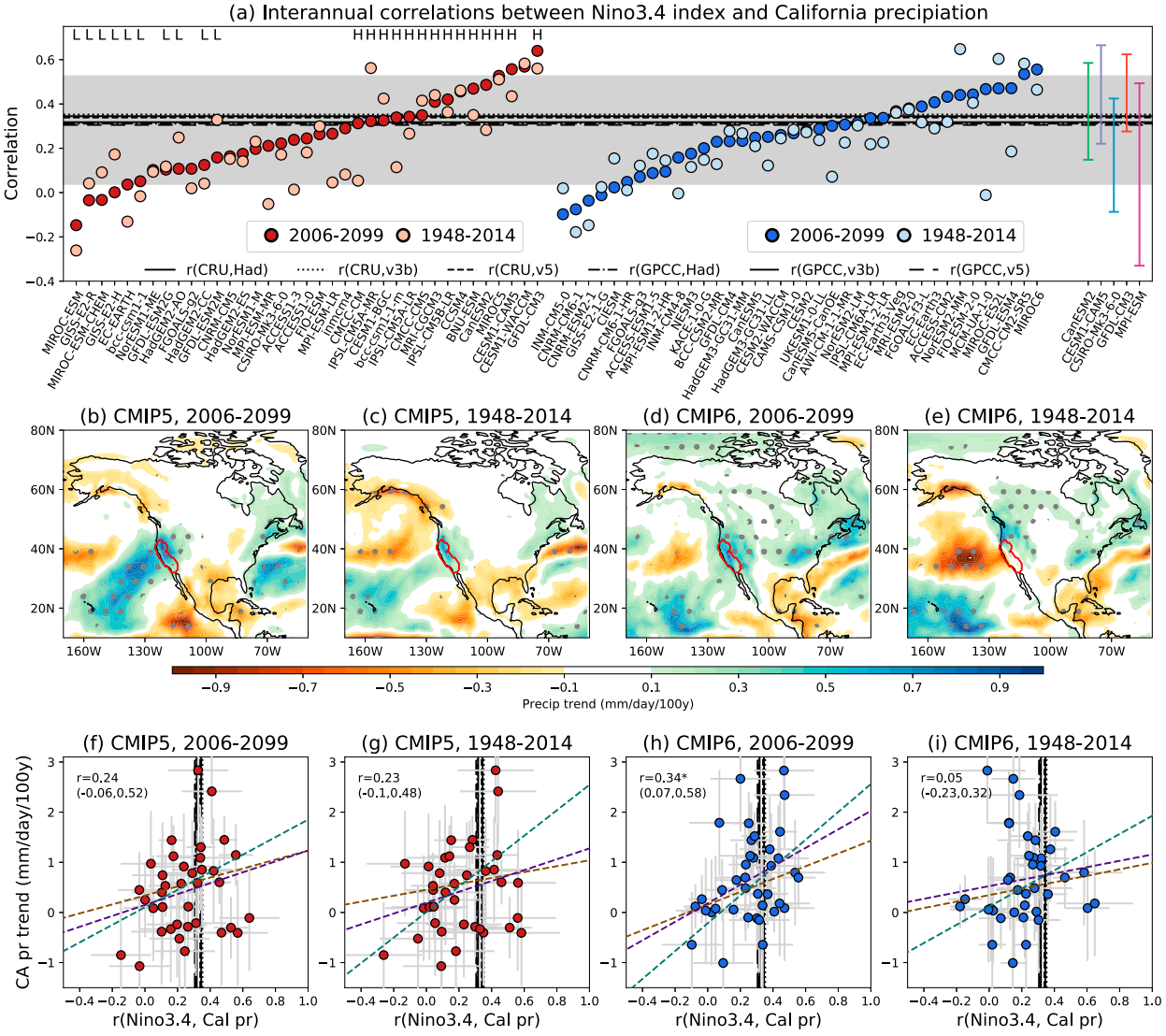

FIG. 13. (a) The interannual correlation between detrended DJF Niño-3.4 and detrended DJF California pr anomalies. Dark red and blue show CMIP5 and CMIP6 values using the 2006–99 period. Light red and light blue show the same but using the 1948–2014 period. Values are ordered according to the 2006–99 correlation, and the H's and L's depicted the models used in the "HIGH-r" and "LOW-r" composites of AL2017. Black lines show the correlation for different combinations of observational SST and precipitation datasets and the gray shading shows the 95% confidence interval on the correlation using GPCC precipitation and ERSSTv5 SSTs. Colored ranges on the right show the minimum to maximum range of the 1948–2014 correlations determined using the LEs. (b)–(e) Differences in 2006–99 precipitation trends between models with a correlation greater than 0.3 and those with a correlation less than 0.2. Stippling indicates regions where the difference is significant at the 95% level using a two-sided bootstrapping test. (f)–(i) The relationship between the 2006–99 precipitation trend over California and the correlation between the Niño-3.4 index and California precipitation. Correlations are quoted along with 95% confidence interval determined by bootstrapping models with replacement. Significant correlations are denoted by an asterisk. Titles in (b)–(i) indicate the time period used to calculate the correlation between the Niño-3.4 index and California precipitation.

overlap with those used in AL2017 (letters H and L in Fig. 13a). We reproduce their result that the ensemble mean of the HIGH-r models shows a relative increase in California precipitation compared to the LOW-r models (Fig. 13b). However, an EC should really be based on a metric determined over the historical period and when using linearly detrended SSTs and precipitation over 1948–2014 the differences between the HIGH-r and LOW-r models is weaker, albeit still positive over California (Fig. 13c). Performing the same calculation in CMIP6 does produce a relative increase in California precipitation in the HIGH-r models when using the 2006–99 period to define *r*(Niño-3.4, pr) (Fig. 13d) but actually produces a relative decrease over much of California in HIGH-r models when using the 1948–2014 period to determine *r*(Niño-3.4, pr) (Fig. 13e).

Figures 13f–i now go beyond the composite difference between the HIGH-r and LOW-r models and show the correlation between California precipitation trends from 2006 to 2099 and r(Niño-3.4, pr). When using the 1948–2014 period, the correlation is not significantly different from zero in either CMIP5 or CMIP6 (Figs. 13g,i) and the same holds true after crudely accounting for model interdependence by first averaging over the models from the same modeling center (supplemental Fig. 6). This will, therefore, not be an effective constraint on California precipitation change, so we do not proceed to quantify it further.

7. Discussion

To perform this analysis, we have proposed a linear regression method that can be used to provide a constraint on future projections while incorporating information on the relevant uncertainties. It is worth discussing how this approach compares to those that have been proposed previously. First, many prior studies that used the linear regression approach have not adequately accounted for the uncertainties involved. Schneider (2018) and Brient (2020) highlight the issues that arise under this circumstance. They demonstrated constrained distributions that were clearly too narrow when compared with model spread. They discussed the various limitations that may lead to this, such as potential inadequacies of the linear model or the OLS fit or disproportionate influences from "bad" models. However, the method they described did not account for the fact that there can be additional spread in the Δy direction that is introduced by internal variability as well as other intermodel differences, not explained by the emergent constraint (our δ term) and, in fact, these are the larger contributors to the uncertainty range (supplemental Fig. 7). Here, we simply introduce these additional uncertainties via sampling procedures and, when doing so, the constrained distributions appear reasonable compared to the spread of Δy values across models that have a predictor aligned with that in observations.

To alleviate concerns related to the inadequacies of the OLS method, we also used TLS and a BHM. We consider the BHM to be the better method given its ability to model not only the uncertainties in $\overline{x}$ and $\Delta\overline{y}$ but also the correlation between them ($r_{x\Delta y}$), as well as to more clearly parse the contributions of each source of uncertainty. However, reassuringly, conclusions are not strongly dependent on the regression method used, although the variance partitioning (Figs. 3 and 8d) and some of the constrained ranges (Fig. 12) do show some small sensitivities. There is perhaps some indication in Figs. 10a–d and 12e–h that the BHM constraint better encompasses the spread of models that have a predictor close to observations. But in each case this is due to one model that lies outside the constrained range, which is not unexpected given that the constrained range is a 95% confidence interval. The overall similarity between methods occurs despite very different approaches in quantifying the uncertainty in the regression coefficients, but is perhaps to be expected since the δ term and the uncertainty due to internal variability in Δy are more important sources of uncertainty than the regression coefficients themselves (supplemental Fig. 7). While we still think it is worthwhile demonstrating the robustness of conclusions across these methods, overall the constraints considered here suggest that the simpler OLS or TLS procedures are adequate, although this may not be the case in situations where the uncertainty in the regression coefficients is relatively more important. The recent study of Tokarska et al. (2020) came to similar conclusions when exploring the sensitivity to methods in the context of constraints on global mean warming.

For each of the methods, it is assumed that a linear relationship exists. This linear relationship may be strongly influenced by models that are highly biased in the predictor and, therefore, may no longer exist once such models are removed from the sample. However, an adequate incorporation of the uncertainties should be able to account for this. Indeed, this does seem to work, as exemplified by the S2016 constraint on U.S. west coast precipitation. For this constraint in CMIP5 there was a significant correlation between the predictor and west coast precipitation (Fig. 11e), but this was clearly being influenced by six strongly biased models (Fig. 11c). When quantifying the constrained distribution of west coast precipitation change, the uncertainties accounted for this, leading to a constrained distribution that was almost as broad as the original model distribution (Fig. 12b).

Another approach to consider is a model weighting procedure (Lorenz et al. 2018; Brient 2020; Brunner et al. 2020), but this has the potential to be strongly influenced by the limited number of models that have a predictor close to observations. If a linear relationship across models does exist, the linear regression method may incorporate more information. We did, however, check that none of our conclusions are qualitatively altered if, instead, the model weighting approach of Brient (2020) is used (not shown).

Recent studies have promoted the use of Bayesian approaches for this problem (Bowman et al. 2018; Williamson and Sansom 2019; Renoult et al. 2020) and our BHM method is strongly aligned with these ideas. It differs in 1) our use of large ensembles to incorporate the uncertainty on modeled and observed values, with dependence on ensemble size in the case of the models, and 2) the resampling procedure that allows for isolation of the different contributions to the uncertainty. In particular, we assess constrained distributions for both the forced change we should expect to occur in the real world (absent internal variability) and the potential future we might experience in the one realization of the real world that we observe (including internal variability).

An implicit assumption when using emergent constraints is that the real world will not behave drastically differently from the model distribution (i.e., it will not deviate from the relationship between predictor and predictand by more than individual models do). This is an assumption that is difficult to test. Williamson and Sansom (2019) describe a method where this additional uncertainty can be incorporated. It remains, however, challenging to quantify what this additional uncertainty should be. Renoult et al. (2020) address this by testing the sensitivity of the constraint to simply inflating the standard deviation of the residuals of their regression fit by a factor of 2. We have not performed such sensitivity tests here because we

have no way of quantifying the problem. So, all of the constraints described above come with the caveat that they assume that the real world is interchangeable with the models in terms of both the relationship between the predictor and the predictand and the magnitude of the additional sources of uncertainty—assumptions that are very difficult to test. For the constraints that were found to agree between CMIP5 and CMIP6, we can at least take comfort in the fact that a different group of models with upgraded physical parameterizations and/or resolution do still obey the same relationships as their predecessors.

8. Conclusions

Three previously proposed emergent constraints have been tested in CMIP6 and a rigorous quantification of the constrained future projections they imply has been provided.

The SHJET constraint (section 4) relates a model's SH wintertime climatological jet position to the magnitude of its future projected poleward shift. This constraint has now been shown to be robust throughout CMIP3, CMIP5 and CMIP6, although in CMIP6 it explains less variance than in CMIP5. Nevertheless, it still provides a quantitatively useful constraint on the future projected poleward shift of the SH westerlies in this season and suggests that there is around an 83% chance that it will be smaller than the mean shift projected by the CMIP5 and CMIP6 ensembles combined (Fig. 5d). However, the mechanism behind this constraint is still not well understood (see the discussion in the online supplemental material).

The VWIND constraint, discussed in section 5, relates a model's response in eddy meridional wind over North America to the amplitude of its climatological, intermediate-scale, stationary waves in that region. This constraint was previously shown to be robust in sensitivity experiments within a single model by van Niekerk et al. (2017) and is also shown here to be robust in CMIP6. However, this is a clear example where model improvements have rendered this constraint less useful. While, on average, the CMIP6 models still have too large a stationary wave amplitude in this region, there are no longer as many models with extremely large biases. As a result, the constraint no longer substantially constrains the future projections beyond the range projected by the CMIP6 models themselves, although it does still suggest a slightly smaller amplitude of the meridional wind change over the U.S. Southwest than the CMIP6 ensemble mean (Fig. 10c).

Extending this to quantify constrained projections on future precipitation change over North America has revealed that, over the U.S. west coast, uncertainties are too large for the relationship between precipitation change and stationary wave amplitude to effectively constrain future projections (Figs. 12a–d). The CMIP5 relationship between stationary wave amplitude and precipitation changes in this region were being dominated by models with very large stationary wave biases and, in CMIP6, now that models have improved, the EC projections are not much more constrained than the CMIP6 distribution itself.

The final constraint we assessed was the CALP constraint, discussed in section 6, which relates a model's representation of the interannual correlation between ENSO and California precipitation r(Niño-3.4, pr), to future trends in California precipitation. In this example we find that there is considerable uncertainty in the predictor r(Niño-3.4, pr) when assessed from the short observational record, which encompasses a large fraction of the model spread (Fig. 13a). Furthermore, we do not find the correlation between historical r(Niño-3.4, pr) and future projected California precipitation change to be robust in either CMIP5 or CMIP6 (Figs. 13g and 13i).

Aside from providing an update on these three ECs, our aim has been to provide a detailed description of methods that can be used to adequately account for uncertainties when constraining future projections, which may now be used to scrutinize other existing and forthcoming ECs. To this end, the analysis codes and methodological descriptions are all provided on github at www.github.com/islasimpson/ecpaper2020/.

Acknowledgments. This project was initiated through a U.S. CLIVAR CMIP6 hackathon. This work is supported by the National Center for Atmospheric Research, which is a major facility sponsored by the National Science Foundation under the Cooperative Agreement 1852977. FL has been supported by the Swiss NSF (Grant PZ00P2_174128) and the Regional and Global Model Analysis (RGMA) component of the Earth and Environmental System Modelling Program of the U.S. Department of Energy's Office of Biological and Environmental Research (BER) vis NSF IA 1844590. We acknowledge the World Climate Research Programme's Working Group on Coupled Modelling, which is responsible for CMIP, and we thank the climate modeling groups (listed in Table 1) for producing and making available their model output. For CMIP, the U.S. Department of Energy's Program for Climate Model Diagnosis and Intercomparison provides coordinating support and led development of software infrastructure in partnership with the Global Organization for Earth System Science Portals. We acknowledge U.S. CLIVAR Working Group on Large Ensembles for curating the large ensemble archive used in this study.

Data availability statement. All climate model data used in this study are available through the CMIP5 and CMIP6 data portals at https://esgf-node.llnl.gov/projects/cmip5 and https://esgf-node.llnl.gov/projects/cmip6. The large ensemble data are available through the U.S. CLIVAR working group on large ensembles Multi-Model Large Ensemble Archive (www.cesm.ucar.edu/projects/community-projects/MMLEA/). ERA5 reanalysis can be downloaded from Copernicus Climate Change services at https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era5; ERA-Interim from https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era-interim; MERRA2 from NASA's GES DISC https://disc.gsfc.nasa.gov/; and JRA-55 from https://jra.kishou.go.jp/JRA-55/index_en.html. GPCC precipitation is available at https://doi.org/10.5676/DWD_GPCC/FD_M_V2018_050 and CRU TS version 4.01 precipitation data are available at https://crudata.uea.ac.uk/cru/data/hrg/cru_ts_4.01/. ERSSTv5 SSTs are available at https://ncei.noaa.gov/pub/data/cmb/ersst/v5/netcdf/, ERSSTv3b at https://www1.ncdc.noaa.gov/pub/data/cmb/ersst/v3b/netcdf/, and HadISST at https://metoffice.gov.uk/hadobs/hadisst/. In addition, all post-processed

data required to reproduce the paper figures can be downloaded from https://doi.org/10.5065/wz6y-5e82 and the processing and analysis scripts are available at www.github.com/islasimpson/ecpaper2020.

APPENDIX

Description of the Bayesian Hierarchical Model Method

In the OLS and TLS approaches, the linear regression model is fit by finding the parameters that minimize the residuals given the data points that we have measured, albeit with different weighting of the errors in the predictor and predictand. The Bayesian hierarchical model (BHM), on the other hand, models the true values $\overline{x}$ and $\Delta\overline{y}$ (uncontaminated by internal variability) based on x and Δy and their uncertainties, allowing for correlation between the errors in x and Δy ($r_{x\Delta y}$) to be incorporated (McKinnon 2015). We assume that δ is represented by a normal distribution with zero mean and variance σ_δ^2 [i.e., $\delta \sim \mathcal{N}(0, \sigma_\delta^2)$], so the regression parameters in the BHM are α, β, and σ_δ^2, and a joint probability distribution of α, β, and σ_δ^2 values is determined, given the values of x and Δy, which we will denote $P(\alpha, \beta, \sigma_\delta^2|x, \Delta y)$. We will continue to use the notation $P(X|Y)$ to denote the conditional probability of X, given Y, throughout.

Bayes' theorem tells us that

$$P(\alpha,\beta,\sigma_\delta^2,\overline{x},\Delta\overline{y}|x,\Delta y) \propto P(x,\Delta y|\alpha,\beta,\sigma_\delta^2,\overline{x},\Delta\overline{y})P(\alpha,\beta,\sigma_\delta^2,\overline{x},\Delta\overline{y}). \tag{A1}$$

Since x and Δy are independent of the regression parameters,

$$P(x,\Delta y|\alpha,\beta,\sigma_\delta^2,\overline{x},\Delta\overline{y}) = P(x,\Delta y|\overline{x},\Delta\overline{y}). \tag{A2}$$

Furthermore, exploiting the identity $P(A, B) = P(A|B)P(B)$, the last term in (A1) can be written as

$$P(\alpha,\beta,\sigma_\delta^2,\overline{x},\Delta\overline{y}) = P(\overline{x},\Delta\overline{y}|\alpha,\beta,\sigma_\delta^2)P(\alpha,\beta,\sigma_\delta^2), \tag{A3}$$

and then exploiting the identity $P(A, B|C) = P(A|B, C)P(B|C)$ gives

$$\begin{aligned} P(\overline{x},\Delta\overline{y}|\alpha,\beta,\sigma_\delta^2) &= P(\Delta\overline{y}|\overline{x},\alpha,\beta,\sigma_\delta^2)P(\overline{x}|\alpha,\beta,\sigma_\delta^2) \\ &= P(\Delta\overline{y}|\overline{x},\alpha,\beta,\sigma_\delta^2)P(\overline{x}) \end{aligned} \tag{A4}$$

and we further assume that $P(\overline{x})$ is best represented by a normal distribution with mean μ_x and standard deviation δ_x^2, such that

$$P(\overline{x}) \propto P(\overline{x}|\mu_x,\delta_x^2)P(\mu_x,\delta_x^2) \tag{A5}$$

while

$$P(\Delta\overline{y}|\overline{x},\alpha,\beta,\sigma_\delta^2) \propto \mathcal{N}(\alpha+\beta\overline{x},\sigma_\delta^2). \tag{A6}$$

So (A1) can be written as

$$\begin{aligned} &P(\alpha,\beta,\sigma_\delta^2,\mu_x,\delta_x^2,\overline{x},\Delta\overline{y}|x,\Delta y) \propto \\ &P(x,\Delta y|\overline{x},\Delta\overline{y})P(\Delta\overline{y}|\overline{x},\alpha,\beta,\sigma_\delta^2)P(\overline{x}|\mu_x\delta_x^2)P(\mu_x,\delta_x^2)P(\alpha,\beta,\sigma_\delta^2), \end{aligned} \tag{A7}$$

which is the joint posterior distribution of all unknowns: $\alpha, \beta, \sigma_\delta^2, x, \Delta y, \mu_x, \delta_x^2$.

Conditional posteriors for $\alpha, \beta, \sigma_\delta^2, \mu_x$, and δ_x^2 can be determined from (A7) by conditioning on all other variables, aside from the one of interest.

We model the relationship between the errors in x and the errors in Δy using a bivariate normal distribution

$$\begin{aligned} P(x,\Delta y) \propto \frac{1}{2\pi\sigma_x,\sigma_{\Delta y}} \exp\Bigg\{ -\frac{1}{2(1-r_{x\Delta y})}\Bigg[&\frac{(x-\overline{x})^2}{\sigma_x^2} \\ &+ \frac{(\Delta y-\Delta\overline{y})^2}{\sigma_{\Delta y}^2} - r_{x\Delta y}\frac{(x-\overline{x})(\Delta y-\Delta\overline{y})}{\sigma_x\sigma_{\Delta y}}\Bigg]\Bigg\} \end{aligned} \tag{A8}$$

which leads to

$$P(x|\Delta y,\overline{x},\Delta\overline{y}) \sim \mathcal{N}\left[x + r_{x\Delta y}\frac{\sigma_x}{\sigma_{\Delta y}}(\Delta y-\Delta\overline{y}), \sigma_x^2\left(1-r_{x\Delta y}^2\right)\right], \tag{A9}$$

$$P(\Delta y|x,\overline{x},\Delta\overline{y}) \sim \mathcal{N}\left[\Delta\overline{y} + r_{x\Delta y}\frac{\sigma_{\Delta y}}{\sigma_x}(x-\overline{x}), \sigma_{\Delta y}^2\left(1-r_{x\Delta y}^2\right)\right], \tag{A10}$$

and use the following priors: $P(\alpha, \beta, \sigma_\delta^2) \propto 1/\sigma_\delta^2$ and $P(\mu_x, \delta_x^2) \propto 1/\delta_x^2$ (i.e., uniform priors on α, β, and μ_x and Jeffrey's prior on σ_δ^2 and δ_x^2). This gives

$$P(\alpha|\cdot) \sim \mathcal{N}\left\{\left[\sum_i \Delta\overline{y}(i) - \beta\sum_i \overline{x}(i)\right]\Big/N, \sigma_\delta^2/N\right\}, \tag{A11}$$

$$P(\beta|\cdot) \sim \mathcal{N}\left\{\frac{\sum_i \overline{x}(i)[\Delta\overline{y}(i)-\alpha]}{\sum_i \overline{x}(i)^2}, \frac{\sigma_\delta^2}{\sum_i \overline{x}(i)^2}\right\}, \tag{A12}$$

$$P(\sigma_\delta^2|\cdot) \sim \mathcal{IG}\left\{N/2, \sum_i [\Delta\overline{y}(i) - \beta\overline{x}(i) - \alpha]^2/2\right\}, \tag{A13}$$

$$P(\mu_x|\cdot) \sim \mathcal{N}\left[\sum_i \overline{x}(i)/N, \sigma_\delta^2/N\right], \tag{A14}$$

$$P(\delta_x^2|\cdot) \sim \mathcal{IG}\left\{N/2, \sum_i [\overline{x}(i) - \mu_x]^2/2\right\}, \tag{A15}$$

where $\mathcal{IG}(A, B)$ refers to an inverse gamma distribution with shape parameter A and scale parameter B and we have used the notation $P(X|\cdot)$ to denote the probability of X conditioned on all other parameters.

Conditional posteriors for $\overline{x}$ and $\Delta\overline{y}$ can be determined from (A7) by conditioning on $\Delta\overline{y}$ and $\overline{x}$, respectively, giving

$$\begin{aligned} &P(\overline{x}|\Delta\overline{y},\alpha,\beta,\sigma_\delta^2,\mu_x,\delta_x^2,x,\Delta y) \propto \mathcal{N}(V_x\Psi_x,\Psi_x), \\ &V_x = \frac{\beta(\Delta\overline{y}-\alpha)}{\sigma_\delta^2} + \frac{x - r_{x\Delta y}\dfrac{\sigma_x}{\sigma_{\Delta y}}(\Delta y-\Delta\overline{y})}{\sigma_x^2(1-r_{x\Delta y}^2)} + \frac{\mu_x}{\delta_x^2}, \\ &\Psi_x = \left[\frac{\beta^2}{\sigma_\delta^2} + \frac{1}{\sigma_x^2(1-r_{x\Delta y}^2)} + \frac{1}{\delta_x^2}\right]^{-1}, \end{aligned} \tag{A16}$$

$$P\left(\Delta\overline{y}|\overline{x},\alpha,\beta,\sigma_\delta^2,\mu_x,\delta_x^2,x,\Delta y\right)\propto\mathcal{N}\left(V_y\Psi_y,\Psi_y\right),$$

$$V_y=\frac{\beta\overline{x}+\alpha}{\sigma_\delta^2}+\frac{\Delta y-r_{x\Delta y}\dfrac{\sigma_{\Delta y}}{\sigma_x}(x-\overline{x})}{\sigma_{\Delta y}^2(1-r_{x\Delta y}^2)},$$

$$\Psi_y=\left[\frac{1}{\sigma_\delta^2}+\frac{1}{\sigma_{\Delta y}^2(1-r_{x\Delta y}^2)}\right]^{-1}. \tag{A17}$$

Thus the full conditional posteriors are represented either by normal distributions or inverse gamma distributions that can be easily sampled. Probability distributions of $\alpha,\beta,\sigma_\delta^2,\mu_x,\delta_x^2,\overline{x}$, and $\Delta\overline{y}$ are obtained using a Markov chain Monte Carlo procedure via Gibbs sampling, whereby a random sample for each of the unknowns is drawn from the conditional posterior distributions [(A11)–(A15) and (A16)–(A17)] in turn. After 30 spinup rounds of sampling, 1000 rounds of samples for each the parameters is performed, giving 1000 ($\alpha,\beta,\sigma_\delta^2,\delta_x^2,\overline{x}$, and $\Delta\overline{y}$) combinations that are then used to provide constrained projections by the methods described in section 3b.

REFERENCES

Allen, R. J., and R. Luptowitz, 2017: El Niño-like teleconnection increases California precipitation in response to warming. *Nat. Commun.*, **8**, 16055, https://doi.org/10.1038/ncomms16055.

Barnes, E. A., and L. Polvani, 2013: Response of the midlatitude jets and of their variability, to increased greenhouse gases in the CMIP5 models. *J. Climate*, **26**, 7117–7135, https://doi.org/10.1175/JCLI-D-12-00536.1.

Boé, J., A. Hall, and X. Qu, 2009: September sea-ice cover in the Arctic Ocean projected to vanish by 2100. *Nat. Geosci.*, **2**, 341–343, https://doi.org/10.1038/ngeo467.

Bowman, K. W., N. Cressie, X. Qu, and A. Hall, 2018: A hierarchical statistical framework for emergent constraints: Application for snow-albedo feedback. *Geophys. Res. Lett.*, **45**, 13 050–13 059, https://doi.org/10.1029/2018GL080082.

Bracegirdle, T. J., and D. B. Stephenson, 2013: On the robustness of emergent constraints used in multimodel climate change projections of Arctic warming. *J. Climate*, **26**, 669–678, https://doi.org/10.1175/JCLI-D-12-00537.1.

——, C. R. Holmes, J. S. Hosking, G. J. Marshall, M. Osman, M. Patterson, and T. Rackow, 2020: Improvements in circumpolar Southern Hemisphere extratropical atmospheric circulation in CMIP6 compared to CMIP5. *Earth Space Sci.*, **7**, e2019EA001065, https://doi.org/10.1029/2019EA001065.

Brient, F., 2020: Reducing uncertainties in climate projections with emergent constraints: Concepts, examples and prospects. *Adv. Atmos. Sci.*, **37** (1), 1–15, https://doi.org/10.1007/s00376-019-9140-8.

Brunner, L., A. G. Pendergrass, F. Lehner, A. L. Merrifield, R. Lorenz, and R. Knutti, 2020: Reduced global warming from CMIP6 projections when weighting models by performance and independence. *Earth Syst. Dyn.*, **11**, 995–1012, https://doi.org/10.5194/esd-11-995-2020.

Caldwell, P. M., C. S. Bretherton, M. D. Zelinka, S. A. Klein, B. D. Santer, and B. M. Sanderson, 2014: Statistical significance of climate sensitivity predictors obtained by data mining. *Geophys. Res. Lett.*, **41**, 1803–1808, https://doi.org/10.1002/2014GL059205.

——, M. D. Zelinka, and S. A. Klein, 2018: Evaluating emergent constraints on equilibrium climate sensitivity. *J. Climate*, **31**, 3921–3941, https://doi.org/10.1175/JCLI-D-17-0631.1.

Chen, X., T. Zhou, P. Wu, Z. Guo, and M. Wang, 2020: Emergent constraints on future projections of the western North Pacific subtropical high. *Nat. Commun.*, **11**, 2802, https://doi.org/10.1038/s41467-020-16631-9.

Cox, P. M., D. Pearson, B. B. Booth, P. Friedlingstein, C. Huntingford, C. D. Jones, and C. M. Luke, 2013: Sensitivity of tropical carbon to climate change constrained by carbon dioxide variability. *Nature*, **494**, 341–344, https://doi.org/10.1038/nature11882.

——, C. Huntingford, and M. S. Williamson, 2018: Emergent constraint on equilibrium climate sensitivity from global temperature variability. *Nature*, **553**, 319–322, https://doi.org/10.1038/nature25450.

Curtis, P. E., P. Ceppi, and G. Zappa, 2020: Role of the mean state for the Southern Hemispheric jet stream response to CO_2 forcing in CMIP6 models. *Environ. Res. Lett.*, **15**, 064011, https://doi.org/10.1088/1748-9326/ab8331.

Dee, D. P., and Coauthors, 2011: The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quart. J. Roy. Meteor. Soc.*, **137**, 553–597, https://doi.org/10.1002/qj.828.

Deser, C., 2020: Certain uncertainty: The role of internal climate variability in projections of regional climate change and risk management. *Earth's Future*, **8**, e2020EF001854, https://doi.org/10.1029/2020EF001854.

——, R. Knutti, S. Solomon, and A. S. Phillips, 2012: Communication of the role of natural variability in future North American climate. *Nat. Climate Change*, **2**, 775–779, https://doi.org/10.1038/nclimate1562.

——, and Coauthors, 2020: Insights from Earth system model initial condition large ensembles and future prospects. *Nat. Climate Change*, **10**, 277–286, https://doi.org/10.1038/s41558-020-0731-2.

Eyring, V., S. Bony, G. A. Meehl, C. A. Senior, B. Stevens, R. J. Stouffer, and K. E. Tayler, 2016: Overview of the Coupled Model Intercomparison Project phase 6 (CMIP6) experimental design and organization. *Geosci. Model Dev.*, **9**, 1937–1958, https://doi.org/10.5194/gmd-9-1937-2016.

Fasullo, J. T., and K. E. Trenberth, 2012: A less cloudy future: The role of subtropical subsidence in climate sensitivity. *Science*, **338**, 792–794, https://doi.org/10.1126/science.1227465.

Gelaro, R., and Coauthors, 2017: The Modern-Era Retrospective Analysis for Research and Applications, version 2 (MERRA-2). *J. Climate*, **30**, 5419–5454, https://doi.org/10.1175/JCLI-D-16-0758.1.

Hall, A., and X. Qu, 2006: Using the current seasonal cycle to constraint snow albedo feedback in future climate change. *Geophys. Res. Lett.*, **33**, L03502, https://doi.org/10.1029/2005GL025127.

——, R. Cox, C. Huntingford, and S. Klein, 2019: Progressing emergent constraints on future climate change. *Nat. Climate Change*, **9**, 269–278, https://doi.org/10.1038/s41558-019-0436-6.

Hargreaves, J. C., J. D. Annan, M. Yoshimori, and A. Abe-Ouchi, 2012: Can the Last Glacial Maximum constrain climate sensitivity? *Geophys. Res. Lett.*, **39**, L24702, https://doi.org/10.1029/2012GL053872.

Harris, I., P. D. Jones, T. J. Osborn, and D. H. Lister, 2014: Updated high-resolution grids of monthly climatic observations—The CRU TS3.10 dataset. *Int. J. Climatol.*, **34**, 623–642, https://doi.org/10.1002/joc.3711.

Hawkins, E., and R. Sutton, 2009: The potential to narrow uncertainty in regional climate predictions. *Bull. Amer. Meteor. Soc.*, **90**, 1095–1108, https://doi.org/10.1175/2009BAMS2607.1.

Hersbach, H., and Coauthors, 2020: The ERA5 global reanalysis. *Quart. J. Roy. Meteor. Soc.*, **146**, 1999–2049, https://doi.org/10.1002/qj.3803.

Huang, B., and Coauthors, 2017: Extended reconstructed sea surface temperature, version 5 (ERSSTv5): Upgrades, validations, and intercomparisons. *J. Climate*, **30**, 8179–8205, https://doi.org/10.1175/JCLI-D-16-0836.1.

Huber, M., I. Mahlstein, M. Wild, J. Fasullo, and R. Knutti, 2010: Constraints on climate sensitivity from radiation patterns in climate models. *J. Climate*, **24** 1034–1052, https://doi.org/10.1175/2010JCLI3403.1.

Kamae, Y., H. Shiogama, M. Watanabe, T. Ogura, T. Yokohata, and M. Kimoto, 2016: Lower-tropospheric mixing as a constraint on cloud feedback in a multiparameter multiphysics ensemble. *J. Climate*, **29**, 6259–6275, https://doi.org/10.1175/JCLI-D-16-0042.1.

Kidston, J., and E. P. Gerber, 2010: Intermodel variability of the poleward shift of the austral jet stream in the CMIP3 integrations linked to biases in 20th century climatology. *Geophys. Res. Lett.*, **37**, L09708, https://doi.org/10.1029/2010GL042873.

Kobayashi, S., and Coauthors, 2015: The JRA-55 reanalysis: General specifications and basic characteristics. *J. Meteor. Soc. Japan*, **93**, 5–48, https://doi.org/10.2151/jmsj.2015-001.

Kriegler, E., and Coauthors, 2017: Fossil-fueled development (SSP5): An energy and resource intensive scenario for the 21st century. *Global Environ. Change*, **42**, 297–315, https://doi.org/10.1016/j.gloenvcha.2016.05.015.

Kwiatkowski, L., L. Bopp, O. Aumont, P. Ciais, P. M. Cox, C. Laufkötter, Y. Li, and R. Séférian, 2017: Emergent constraints on projections of declining primary production in the tropical oceans. *Nat. Climate Change*, **7**, 355–358, https://doi.org/10.1038/nclimate3265.

Lamarque, J. F., G. P. Kyle, M. Meinshausen, K. Riahi, S. J. Smith, D. P. van Vuuren, A. J. Conley, and F. Vitt, 2011: Global and regional evolution of short-lived radiatively active gases and aerosols in the representative concentration pathways. *Climatic Change*, **109**, 191–212, https://doi.org/10.1007/s10584-011-0155-0.

Lehner, F., A. W. Wood, J. A. Vano, D. M. Lawrence, M. P. Clark, and J. S. Makin, 2019: The potential to reduce uncertainty in regional runoff projections from climate models. *Nat. Climate Change*, **9**, 926–933, https://doi.org/10.1038/s41558-019-0639-x.

——, C. Deser, N. Maher, J. Marotzke, E. Fischer, L. Brunner, R. Knutti, and E. Hawkins, 2020: Partitioning climate projection uncertainty with multiple large ensembles and CMIP5/6. *Earth Syst. Dyn.*, **11**, 491–508, https://doi.org/10.5194/esd-11-491-2020.

Li, G., S.-P. Xie, C. He, and Z. S. Chen, 2017: Western Pacific emergent constraint lowers projected increase in Indian summer monsoon rainfall. *Nat. Climate Change*, **7**, 708–712, https://doi.org/10.1038/nclimate3387.

Lipat, B. R., G. Tselioudis, K. M. Grise, and L. M. Polvani, 2017: CMIP5 models' shortwave cloud radiative response and climate sensitivity linked to the climatological Hadley cell extent. *Geophys. Res. Lett.*, **44**, 5739–5748, https://doi.org/10.1002/2017GL073151.

Lorenz, R., N. Herger, J. Sedlacek, V. Eyring, E. M. Fischer, and R. Knutti, 2018: Prospects and caveats for weighting climate models for summer maximum temperature projections over North America. *J. Geophys. Res. Atmos.*, **123**, 4509–4526, https://doi.org/10.1029/2017JD027992.

Massonnet, F., T. Fichefet, H. Goosse, C. M. Bitz, G. Philippon-Berthier, M. M. Holland, and P.-Y. Barriat, 2012: Constraining projections of summer Arctic sea ice. *Cryosphere*, **6**, 1383–1394, https://doi.org/10.5194/tc-6-1383-2012.

McKinnon, K. A., 2015: Understanding and predicting temperature variability in the observational record. Ph.D. thesis, Harvard University, 143 pp., https://dash.harvard.edu/handle/1/17463140.

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

O'Gorman, P. A., 2012: Sensitivity of tropical precipitation extremes to climate change. *Nat. Geosci.*, **5**, 697–700, https://doi.org/10.1038/ngeo1568.

O'Neill, B. C. O., E. Kriegler, K. Riahi, K. L. Ebi, S. Hallegatte, T. R. Carter, R. Mathur, and D. P. van Vuuren, 2013: A new scenario framework for climate change research: The concept of shared socioeconomic pathways. *Climatic Change*, **122**, 387–400, https://doi.org/10.1007/s10584-013-0905-2.

Rayner, N. A., D. E. Parker, E. B. Horton, C. K. Folland, L. V. Alexander, D. P. Rowell, E. C. Kent, and A. Kaplan, 2003: Global analyses of sea surface temperature, sea ice and night marine air temperature since the late nineteenth century. *J. Geophys. Res.*, **108**, 4407, https://doi.org/10.1029/2002JD002670.

Renoult, M., and Coauthors, 2020: A Bayesian framework for emergent constraints: Case studies of climate sensitivity with PMIP. *Climate Past Discuss.*, **16**, 1715–1735, https://doi.org/10.5194/cp-16-1715-2020.

Schneider, T., 2018: Statistical inference with emergent constraints. Caltech Climate Dynamics Group blog, 24 January 2018, https://climate-dynamics.org/statistical-inference-with-emergent-constraints/.

Shepherd, T. G., and Coauthors, 2018: Storylines: An alternative approach to representing uncertainty in physical aspects of climate change. *Climatic Change*, **151**, 555–571, https://doi.org/10.1007/s10584-018-2317-9.

Sherwood, S. C., S. Bony, and J.-L. Dufresne, 2014: Spread in model climate sensitivity traced to atmospheric convective mixing. *Nature*, **505**, 37–42, https://doi.org/10.1038/nature12829.

Simpson, I. R., and L. M. Polvani, 2016: Revisiting the relationship between jet position, forced response, and annular mode variability in the southern midlatitudes. *Geophys. Res. Lett.*, **43**, 2896–2903, https://doi.org/10.1002/2016GL067989.

——, R. Seager, M. Ting, and T. A. Shaw, 2016: Causes of change in Northern Hemisphere winter meridional winds and regional hydroclimate. *Nat. Climate Change*, **6**, 65–70, https://doi.org/10.1038/nclimate2783.

Smith, T. M., R. W. Reynolds, T. C. Peterson, and J. Lawrimore, 2008: Improvements to NOAA's historical merged land–ocean surface temperature analysis (1880–2006). *J. Climate*, **21**, 2283–2296, https://doi.org/10.1175/2007JCLI2100.1.

Son, S., and Coauthors, 2010: Impact of stratospheric ozone on the Southern Hemisphere circulation changes: A multimodel assessment. *J. Geophys. Res. Atmos.*, **115**, 1–55, https://doi.org/10.1029/2010JD014271.

Su, H., J. H. Jiang, C. Zhai, T. Shen, J. D. Neelin, G. L. Stephens, and Y. L. Yung, 2014: Weakening and strengthening structures in the Hadley circulation change under global warming and implications for cloud response and climate sensitivity. *J. Geophys. Res.*, **119**, 5787–5805, https://doi.org/10.1002/2014JD021642.

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.

Tian, B., 2015: Spread of model climate sensitivity linked to double-intertropical convergence zone bias. *Geophys. Res. Lett.*, **42**, 4133–4141, https://doi.org/10.1002/2015GL064119.

Tokarska, K. B., M. B. Stope, S. Sippel, E. M. Fischer, C. J. Smith, F. Lehner, and R. Knutti, 2020: Past warming trend constraints future warming in CMIP6 models. *Sci. Adv.*, **6**, eaaz9549, https://doi.org/10.1126/sciadv.aaz9549.

Trenberth, K. E., and J. T. Fasullo, 2010: Simulation of present-day and twenty-first-century energy budgets of the southern oceans. *J. Climate*, **23**, 440–454, https://doi.org/10.1175/2009JCLI3152.1.

van Niekerk, A., J. F. Scinocca, and T. G. Shepherd, 2017: The modulation of stationary waves, and their response to climate change, by parameterized orographic draft. *J. Atmos. Sci.*, **74**, 2557–2574, https://doi.org/10.1175/JAS-D-17-0085.1.

Volodin, E. M., 2008: Relation between temperature sensitivity to doubled carbon dioxide and the distribution of clouds in current climate models. *Izv. Atmos. Oceanogr. Phys.*, **44**, 288–299, https://doi.org/10.1134/S0001433808030043.

Wagman, B. M., and C. S. Jackson, 2018: A test of emergent constraints on cloud feedback and climate sensitivity using a calibrated single model ensemble. *J. Climate*, **31**, 7515–7532, https://doi.org/10.1175/JCLI-D-17-0682.1.

Wenzel, S., P. M. Cox, V. Eyring, and P. Friedlingstein, 2014: Emergent constraints on climate-carbon cycle feedbacks in the CMIP5 Earth system models. *J. Geophys. Res. Biogeosci.*, **119**, 794–807, https://doi.org/10.1002/2013JG002591.

Williamson, D. B., and P. G. Sansom, 2019: How are emergent constraints quantifying uncertainty and what do they leave behind? *Bull. Amer. Meteor. Soc.*, **100**, 2571–2588, https://doi.org/10.1175/BAMS-D-19-0131.1.

Zhai, C., J. H. Jiang, and H. Su, 2015: Long-term cloud change imprinted in seasonal cloud variation: More evidence of high climate sensitivity. *Geophys. Res. Lett.*, **42**, 8729–8737, https://doi.org/10.1002/2015GL065911.