SUPPLEMENTARY MATERIAL

1 Comparison with existing methods

1.1 Ensemble regression

Bracegirdle and Stephenson (2012) proposed a method for projection using emergent constraints known as "ensemble regression". Ensemble regression is equivalent to simple linear regression of the model mean responses on the model mean historical climates, and can be written in our notation as

$$\bar{X}_{Fm} - \bar{X}_{Hm} \sim N\left(\bar{X}_F - \bar{X}_H + \beta'(\bar{X}_{Hm} - \bar{X}_H), \sigma^2_{F|H}\right)$$

where $\bar{X}_{tm} = \sum_r X_{tmr}/R_{tm}$ and $\bar{X}_t = \sum_m \bar{X}_{tm}/M$. This is equivalent to our Equation 2 in the main text where $\beta' = \beta - 1$, since $\mathrm{E}\left[\bar{X}_{tm}\right] = X_{tm}$ and $\mathrm{E}\left[\bar{X}_t\right] = \mu_t$.

Ensemble regression ignores uncertainty due to internal variability in the model means $\bar{X}_{Hm}$ and the ensemble mean $\bar{X}_H$. It is well known that errors in the independent variable $(\bar{X}_{Hm} - \bar{X}_H)$ in a regression will cause the slope estimate to be biased towards zero, a phenomenon known as *regression dilution* or *regression attenuation* (Frost and Thompson, 2000). Consider a balanced ensemble ($R_{Hm} = R_{Fm} = R$ for all m) in which all models simulate the same internal variability in each time period, i.e., $\sigma^2_m = \sigma^2$ and $\varphi^2_m = 1$ for all m. The expected value of the linear regression estimate of the emergent constraint is

$$\mathrm{E}\left[\hat{\beta}'\right] = \frac{\mathrm{cov}\left(\bar{X}_{Fm} - \bar{X}_{Hm}, \bar{X}_{Hm} - \bar{X}_H\right)}{\mathrm{var}\left(\bar{X}_{Hm} - \bar{X}_H\right)} = \frac{\beta'\sigma^2_H - \sigma^2/R}{\sigma^2_H + \sigma^2/R}$$

where β' is the "true" value of the emergent constraint. The bias is largest when the internal variability σ^2 is large compared to the model uncertainty σ^2_H, or when the number

of runs R from each model is small. Our framework avoids this bias by explicitly modeling internal variability and its relationship to the expected model climates X_{tm}.

In Bracegirdle and Stephenson (2012), the ensemble regression estimate of the response of the Earth system is

$$Y_F - Y_H \sim N\left(\bar{X}_F - \bar{X}_H + \beta'(Z_H - \bar{X}_H), \sigma^2_{F|H}\right).$$

This is equivalent to assuming the Earth system is exchangeable with the models and ignores the possibility of common differences between the models and the Earth system, as well as the effects of observation uncertainty and natural variability. The framework proposed here explicitly allows for common model inadequacy, observation uncertainty and natural variability.

1.2 A simple hierarchical framework

Bowman et al. (2018) propose a hierarchical framework for emergent constraints without explicit reference to climate models. In our notation, the linear normal-theory version is

$$Y_H \sim N\left(\mu_H, \sigma^2_H\right) \qquad Y_F \mid Y_H \sim N\left(\mu_F + \beta(Y_H - \mu_H), \sigma^2_{F|H}\right)$$

and $Z_H \mid Y_H \sim N(Y_H, \sigma^2_Z)$. In practice, the parameters μ_H, μ_F, β, σ_H and $\sigma_{F|H}$ are estimated from an ensemble of climate models by assuming

$$X_{Hm} \sim N\left(\mu_H, \sigma^2_H\right) \qquad X_{Fm} \mid X_{Hm} \sim N\left(\mu_F + \beta(X_{Hm} - \mu_H), \sigma^2_{F|H}\right)$$

for all $m = 1, \ldots, M$. This is identical to Equation 2 in the main text, so the framework proposed by Bowman et al. (2018) is almost equivalent to Ensemble Regression (Bracegirdle and Stephenson, 2012), but allowing for observation uncertainty. However, the inclusion

of the prior on Y_H implies that the posterior expectation of Y_F (Bowman et al., 2018, Eqns 11,13,17) is

$$\mathrm{E}\left[Y_F \mid Z_H\right] = \mu_F + \beta \frac{\sigma_Z^{-2}}{\sigma_Z^{-2} + \sigma_H^{-2}} \left(Z_H - \mu_H\right).$$

So the expected future climate Y_F experiences a shrinkage towards the representative climate μ_F depending on how informative the observations are compared to the models for the historical climate Y_H, i.e., the ratio of σ_Z^2 to σ_H^2. No attempt is made to account for model inadequacy, the Earth system is implicitly assumed to be exchangeable with the models. Further, only one run from each model is used, thus ignoring internal variability and leaving the estimated emergent relationship vulnerable to regression dilution.

1.3 The coexchangeable framework

Rougier et al. (2013) propose a model of the joint distribution of the historical and future climate in multi-model experiments known as the coexchangeable framework. In our notation

$$\begin{aligned} \mathbf{X}_m &\sim N\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) & m &= 1, \ldots, M \\ \mathbf{Y} &\sim N\left(\mathbf{A}\boldsymbol{\mu}, \boldsymbol{\Sigma}_\Delta\right) & Z_H &\sim N\left(Y_H, \sigma_Z^2\right) \end{aligned}$$

where $\mathbf{X}_m = (X_{Hm}, X_{Fm})^T$, $\mathbf{Y} = (Y_H, Y_F)^T$, $\boldsymbol{\mu} = (\mu_H, \mu_F)^T$. The matrix $\mathbf{A}$ is assumed known and allows for transformation of variables between model world and the real world (the default choice is $\mathbf{A} = \mathbf{I}$, the identity). The exchangeable framework is a special case of the the coexchangeable framework where $\boldsymbol{\Sigma}_\Delta = \boldsymbol{\Sigma}$ and $\mathbf{A} = \mathbf{I}$. The framework proposed here is an extension of the coexchangeable framework with $\mathbf{A} = \mathbf{I}$ and

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_H \\ \mu_F \end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_H^2 & \beta\sigma_H^2 \\ \beta\sigma_H^2 & \beta^2\sigma_H^2 + \sigma_{F|H}^2 \end{pmatrix}.$$

However, the basic coexchangeable framework does not distinguish between model differences and internal variability, and does not account for natural variability in the Earth system. The extended framework proposed here accounts for both of these additional sources of uncertainty.

Rougier et al. (2013) suggest the following parametrization of the model inadequacy

$$\mathbf{\Sigma}_\Delta = \kappa^2 \mathbf{\Sigma} + \mathbf{D}$$

where $\mathbf{D}$ is a diagonal matrix with $diag(\mathbf{D}) = (D_H^2, D_F^2)^T$. The variances D_H^2 and D_F^2 are intended to guard against overly precise projections when models are in close agreement. However, this parametrization has unexpected consequences for emergent constraints. Standard results for the multivariate normal distribution show that

$$\mathrm{E}\left[Y_F \mid Y_H\right] = \mu_F + \beta^\star (Y_H - \mu_H) \quad \text{where} \quad \beta^\star = \frac{\mathrm{cov}\left(Y_F, Y_H\right)}{\mathrm{var}\left(Y_H\right)} = \frac{\kappa^2 \sigma_H^2}{\kappa^2 \sigma_H^2 + D_H^2} \beta$$

The emergent constraint shrinks towards zero by an amount that depends on D_H^2. This is difficult to defend given that we have assumed the emergent constraint has a physical basis and should apply to the Earth system. Similar terms D_H^2 and $D_{F|H}^2$ could be added to Equation 8 in the main text, but without effecting the emergent constraint, since then $\mathrm{cov}\left(Y_F, Y_H\right) = \mathrm{var}\left(Y_H\right) = \kappa^2 \sigma_H^2 + D_H^2$ and $\beta^\star = \beta$. The difference is due to our formulation in terms of conditional rather than marginal variances. Like $\sigma_{\Delta_H}^2$ and $\sigma_{\Delta_{F|H}}^2$, D_H^2 and $D_{F|H}^2$ are difficult to specify *a priori* without additional data. One possibility might be to consider the spread of a family of closely related models as a lower bound for the model inadequacy.

1.4 The generalized truth-plus-error framework

Chandler (2013) proposed an alternative joint framework for multi-model projection

$$\begin{aligned} \mathbf{X}_{mr} &\sim N\left(\mathbf{X}_m, \mathbf{\Sigma}_m\right) & r &= 1, \ldots, N_m \\ \mathbf{X}_m &\sim N\left(\boldsymbol{\mu}, \mathbf{\Lambda}_m\right) & m &= 1, \ldots, M \\ \boldsymbol{\mu} &\sim N\left(\mathbf{Y}, \mathbf{\Sigma}_\Delta\right) & Y_{Ha} &\sim N\left(Y_H, \sigma_a^2\right) \end{aligned}$$

where $\mathbf{X}_{mr} = (X_{Hmr}, X_{Fmr})^T$, $\mathbf{X}_m = (X_{Hm}, X_{Fm})^T$, $\boldsymbol{\mu} = (\mu_H, \mu_F)^T$, $\mathbf{Y} = (Y_H, Y_F)^T$. The variances $\mathbf{\Lambda}_{tm}$ represent the propensity of each simulator to deviate from the ensemble consensus. This provides flexibility to incorporate prior knowledge that certain climate models are more or less similar to each other. Internal variability and model inadequacy are both accounted for. In contrast to Rougier et al. (2013), natural variability is accounted for, but observation uncertainty is ignored.

Chandler (2013) suggests estimating the historical model inadequacy from data as $\sigma^2_{\Delta_H} = (Y_{Ha} - \mu_H)^2$ then setting

$$\mathbf{\Sigma}_\Delta = \begin{pmatrix} \sigma^2_{\Delta_H} & \sigma^2_{\Delta_H} \\ \sigma^2_{\Delta_H} & (1+\kappa)\sigma^2_{\Delta_H} \end{pmatrix}$$

for $\kappa > 0$. This parametrization ignores any emergent constraints in the projection of the future climate. In addition, estimating $\sigma^2_{\Delta_H}$ from a single observation Y_{Ha} provides very limited information and makes the analysis vulnerable to outlying or spurious measurements.

The frameworks proposed by Chandler (2013) and Rougier et al. (2013) are conceptually very different and appear incompatible. The most obvious difference is the direction of conditioning between the system $\mathbf{Y}$ and the representative or consensus climate $\boldsymbol{\mu}$. However, Rougier et al. (2013) demonstrated that a simplified form of the generalized truth-plus-error framework can be viewed as a special case of the coexchangeable framework (up to

the second moments), for particular choices of $\mathbf{A} \neq \mathbf{I}$ and $\mathbf{\Sigma}_\Delta$. It is interesting to note that when all the distributions are normal, identical priors are set for related quantities and $\mathbf{A} = \mathbf{I}$, both frameworks produce identical posterior inferences. This is not the case when the assumption of normality is relaxed, and should not be interpreted as meaning that both formulations are equivalent and can be used interchangeably.

? also considered the direction of conditioning between climate models the Earth system and concluded that it should be decided by our ability to formulate the relevant distributions, to interpret them, and to perform the necessary computations. We find it more natural to consider the actual climate as the sum of our knowledge (the representative model) plus what we do *not* understand (model inadequacy), than vice-versa. Hence we adopt a coexchangeable representation for the models. In contrast, in the supplementary material we use a truth-plus-error representation to combine reanalysis data sets in order to estimate observation uncertainty. This feels more natural since the reanalyses are trying to approximate an observable (Y_{Ha}), rather than an abstract quantity ("the climate") for the models.

1.5 Reliability ensemble averaging

Tebaldi et al. (2005) proposed a probabilistic interpretation of the heuristic "reliability ensemble averaging" framework of **?**. The framework belongs to the truth-plus-error family, with some interesting features. Multivariate extensions were proposed by Smith et al. (2009) and Tebaldi and Sansó (2009), and a similar spatial framework was proposed by Furrer et al. (2007). The basic framework in our notation is given by

$$X_{Hm} \sim N\left(Y_H, \lambda_m^2\right) \qquad X_{Fm} \mid X_{Hm} \sim N\left(Y_F + \beta(X_{Hm} - Y_H), (\theta\lambda_m)^2\right)$$
$$Y_{Ha} \sim N\left(Y_H, \sigma_a^2\right).$$

Similar to Chandler (2013), the model climates X_{tm} are conditioned on the Earth system climate Y_t, and the variances λ_m are interpreted as the propensity of each model to deviate from the system. The coefficient θ allows the propensity of the models to differ from the system to change in the future period. Somewhat confusingly, natural variability in the Earth system is accounted for, but internal variability in the models is ignored. Observation uncertainty and model inadequacy are also both neglected.

The framework proposed by Tebaldi et al. (2005) includes something similar to an emergent constraint. It is instructive to consider this alternative formulation in detail. The expectation of the full conditional posterior distribution of future climate (Tebaldi et al., 2005, Eqn. A9) is

$$\mathrm{E}\left[Y_F \mid \ldots\right] = \frac{\sum_m \lambda_m^{-2} X_{Fm}}{\sum_m \lambda_m^{-2}} + \beta \left(Y_H - \frac{\sum_m \lambda_m^{-2} X_{Hm}}{\sum_m \lambda_m^{-2}} \right).$$

This is equivalent to Equation 5 in the main text, if $\lambda_m^2 = \sigma_H^2$ for all m, i.e., if all the models are exchangeable. Let $\lambda_m^2 = \sigma_H^2$ and $\theta\lambda_m^2 = \sigma_{F|H}^2$ for all m, then the posterior expectation of Y_H (Tebaldi et al., 2005, Eqn. A8) is

$$\mathrm{E}\left[Y_H \mid \ldots\right] = \frac{\sigma_a^{-2} Y_{Ha} + M\left(\sigma_H^{-2}\bar{X}_H + \sigma_{F|H}^{-2}\beta(Y_F - \bar{X}_F + \beta\bar{X}_H)\right)}{\sigma_a^{-2} + M\left(\sigma_H^{-2} + M\sigma_{F|H}^{-2}\beta^2\right)}$$

In comparison, the posterior expectation of Y_H in our framework is

$$\mathrm{E}\left[Y_H \mid \ldots\right] = \frac{\sigma_a^{-2} Y_{Ha} + \sigma_H^{-2}\mu_H + \sigma_{F|H}^{-2}\beta\left(Y_F - \mu_F + \beta\mu_H\right)}{\sigma_a^{-2} + \sigma_H^{-2} + \sigma_{F|H}^{-2}\beta^2}$$

assuming $\kappa = 1$, i.e., the models are exchangeable with the Earth system. Both estimates are weighted averages of the model outputs and the actualized climate Y_{Ha}. The two estimates effectively differ only in the weight given to the models. Under the framework proposed by Tebaldi et al. (2005), the models receive M times more weight than under

our framework. As a result, the posterior expectation of the expected climate Y_H, and consequently the projected climate Y_F, will lie much closer to the consensus climate, and approach the consensus as the number of models increases. In fact, the framework proposed by Tebaldi et al. (2005) implies that we can learn the expected climate Y_H and Y_F to any degree of precision we require, simply by adding more climate models (**?**). Given the existence of shared errors in all climate models, such an assumption is unsupportable. Tebaldi and Sansó (2009) later proposed the inclusion of a common model bias term to address this issue. However, the common bias was treated as a fixed quantity to be estimated, and does not contribute to our uncertainty about the Earth system in the same way as the model inadequacy terms proposed here and by Rougier et al. (2013) and Chandler (2013).

1.6 Model weighting and Bayesian model averaging

A variety of model weighting schemes have been proposed in the literature (the Introduction in the main text for examples), but all have essentially the same functional form

$$Y_t = \sum_{m=1}^{M} w_m X_{tm}.$$

The actual climate (or climate response) Y_t is modeled as a weighted combination of the model outputs. Depending on the exact formulation, the weights w_m may be constrained to be positive and sum to one. The weights w_m are estimated by comparing observations of the historical climate Y_H with model simulations X_{Hm} of the same period. The same weights are then applied to future simulations X_{Fm} to obtain projections of the future climate Y_F or climate response $Y_F - Y_H$. Bayesian Model Averaging differs from simple model weighting by dressing each simulation X_{tm} with a kernel, so Y_t becomes a mixture model.

In principle, model weighting will respect emergent relationships. Consider the example of Figure 1 in the main text. If the models closest to the observations receive the most weight, then the projected climate response will be lower than the ensemble mean estimate. However, unless the models further from the observations receive almost zero weight, the projected response will shrink towards the ensemble mean. The amount of shrinkage will depend on the exact form of the weights. In practice, the weights w_m are usually estimated by comparing model performance at multiple locations, often across the entire study region (e.g., Bhat et al., 2011; Knutti et al., 2017). If the emergent relationship does not apply across the entire region, or varies within the region, then the weights are unlikely to reflect the relationship and the constraining behavior will be lost.

2 Ensemble thinning

An extended version of the CMIP5 surface temperature data analyzed by Bracegirdle and Stephenson (2013) was considered for analysis. The mean climates over 30 winters (December-January-February) are compared between December 1975 and January 2005 from the historical scenario, and between December 2069 and January 2099 from the RCP4.5 scenario. The five year shift in the historical period compared to Bracegirdle and Stephenson (2013) provides slightly better compatibility with the latest observation and reanalysis data sets. Several of these data sets begin in 1979 when satellite observations become prevalent. A total of 216 runs from 37 CMIP5 models were included in the full ensemble, 128 runs of the historical scenario and 88 of the RCP4.5 scenario. The complete list of models and details of their major components are given in Table 1.

In the main text we noted that not all of the models should be included in the analysis in order to satisfy the assumption of exchangeability. In particular, models from the same center are likely to be more similar than those from different centers. Therefore, only one model from each center should be included. Modeling centers may also share components with other groups. Therefore, where possible only one model using any given major component, or at least any combination of components, should be included.

The full ensemble was thinned in order to satisfy the judgment of exchangeability between the model outputs. The ACCESS models supersede the CSIRO-Mk3.6.0 model, however all of the major components in the ACCESS models are borrowed from other models. Therefore, none of the models submitted by CSIRO were included. Two models were submitted by BCC, the model with the higher resolution atmosphere components was retained. Three models were submitted from the combined efforts of the NSF-DOE-NCAR. The CESM1(CAM5) variant was selected as it includes a more recent version of the CAM

Modelling centre	Model	Atmosphere	Atmosphere res.	Ocean	Ocean res.	Sea ice	Land surface
CSIRO-BOM	ACCESS1.0	HadGEM2 (r1.1)	1.9 x 1.9 L38	MOM4.1	1.0 x 1.0 L50	CICE4.1	MOSES2.2
CSIRO-BOM	ACCESS1.3	UKMO GA 1.0	1.9 x 1.9 L38	MOM4.1	1.0 x 1.0 L50	CICE4.1	CABLE
BCC	BCC-CSM1.1	BCC_AGCM2.1	2.8 x 2.8 L26	MOM4-L40	1.0 x 1.0 L40	GFDL SIS	BCC-AVIM1.0
BCC	BCC-CSM1.1(m)	BCC_AGCM2.1	1.1 x 1.1 L26	MOM4-L40	1.0 x 1.0 L40	GFDL SIS	BCC-AVIM1.0
GCESS	BNU-ESM	CAM3.5	2.8 x 2.8 L26	MOM4.1	1.0 x 1.0 L50	CICE4.1	CoLM+BNUDGVM
CCCMA	CanESM2	CanAM4	2.8 x 2.8 L35	CanOM4	1.4 x 1.4 L40	Included	CLASS 2.7; CTEM
NCAR	CCSM4	CAM4	1.2 x 1.2 L27	POP2	1.0 x 1.0 L60	CICE4	CLM4
NSF-DOE-NCAR	CESM1(BGC)	CAM4	1.2 x 1.2 L27	POP2	1.0 x 1.0 L60	CICE4	CLM4
NSF-DOE-NCAR	CESM1(CAM5)	CAM5	1.2 x 1.2 L27	POP2	1.0 x 1.0 L60	CICE4	CLM4
NSF-DOE-NCAR	CESM1(WACCM)	WACCM4	2.5 x 2.5 L66	POP2	1.0 x 1.0 L60	CICE4	CLM4
CMCC	CMCC-CM	ECHAM5	0.8 x 0.8 L31	OPA8.2	2.0 x 2.0 L31	LIM2	N / A
CMCC	CMCC-CMS	ECHAM5	1.9 x 1.9 L95	OPA8.2	2.0 x 2.0 L31	LIM2	N / A
CNRM-CERFACS	CNRM-CM5	ARPEGE-Climat	1.4 x 1.4 L31	NEMO	0.7 x 0.7 L42	Gelato5	SURFEX
CSIRO-QCCCE	CSIRO-Mk3.6.0	Included	1.9 x 1.9 L18	MOM2.2	1.9 x 1.9 L31	Included	Included
EC-EARTH	EC-EARTH	IFS c3 1r1	1.1 x 1.1 L62	NEMO_ecmwf	1.0 x 1.0 L31	LIM2	HTESSEL
LASG-CESS	FGOALS-g2	GAMIL2	2.8 x 2.8 L26	LICOM2	1.0 x 1.0 L30	CICE4-LASG	CLM3
FIO	FIO-ESM	CAM3.0	2.8 x 2.8 L26	POP2	1.0 x 1.0 L40	CICE4	CLM3.5
NOAA GFDL	GFDL-CM3	Included	2.5 x 2.5 L48	MOM4.1	1.0 x 1.0 L50	SIS	Included
NOAA GFDL	GFDL-ESM2G	AM2.1	2.5 x 2.5 L24	GOLD	1.0 x 1.0 L63	SIS	Included
NOAA GFDL	GFDL-ESM2M	AM2.1	2.5 x 2.5 L60	MOM4.1	1.0 x 1.0 L50	SIS	Included
NASA GISS	GISS-E2-H	Inlcuded	2.5 x 2.5 L40	HYCOM	1.0 x 1.0 L26	Included	Included
NASA GISS	GISS-E2-R	Inlcuded	2.5 x 2.5 L40	Russell	1.0 x 1.0 L32	Included	Included
NIMR/KMA	HadGEM2-AO	HadGAM2	1.9 x 1.9 L60	Included	1.9 x 1.9	Included	Included
MOHC	HadGEM2-CC	HadGAM2	1.9 x 1.9 L60	Included	1.9 x 1.9	Included	Included
MOHC	HadGEM2-ES	HadGAM2	1.9 x 1.9 L38	Included	1.0 x 1.0 L40	Included	Included
INM	INM-CM4	Included	2.0 x 2.0 L21	Included	1.0 x 1.0 L40	Included	Included
IPSL	IPSL-CM5A-LR	LMDZ5	3.8 x 3.8 L39	NEMO	2.0 x 2.0 L31	Included	Included
IPSL	IPSL-CM5A-MR	LMDZ5	2.5 x 2.5 L39	NEMO	2.0 x 2.0 L31	Included	Included
IPSL	IPSL-CM5B-LR	LMDZ5	3.8 x 3.8 L39	NEMO	2.0 x 2.0 L31	Included	Included
MIROC	MIROC5	MIROC-AGCM6	1.4 x 1.4 L40	COCO4.5	1.4 x 1.4 L50	Included	MATSIRO
MIROC	MIROC-ESM	MIROC-AGCM	2.8 x 2.8 L80	COCO3.4	1.4 x 1.4 L44	Included	MATSIRO
MIROC	MIROC-ESM-CHEM	MIROC-AGCM	2.8 x 2.8 L80	COCO3.4	1.4 x 1.4 L44	Included	MATSIRO
MPI-M	MPI-ESM-LR	ECHAM6	1.9 x 1.9 L47	MPIOM	1.5 x 1.5 L40	Included	JSBACH
MPI-M	MPI-ESM-MR	ECHAM6	1.9 x 1.9 L95	MPIOM	0.4 x 0.4 L40	Included	JSBACH
MRI	MRI-CGCM3	MRI-AGCM3.3	1.1 x 1.1 L48	MRI.COM3	1.0 x 1.0 L50	MRI.COM3	HAL
NCC	NorESM1-M	CAM4-Oslo	2.5 x 2.5 L26	NorESM-Ocean	1.1 x 1.1 L53	CICE4	CLM4
NCC	NorESM1-ME	CAM4-Oslo	2.5 x 2.5 L26	NorESM-Ocean	1.1 x 1.1 L53	CICE4	CLM4

Table 1: Structural details of the 37 CMIP5 models considered for the analysis of Arctic surface temperature. Models highlighted in red are included in the exchangeable ensemble. Atmosphere and ocean resolution are in degrees and Lxx indicates the number of vertical levels. Details included in this table were gathered from the metadata included in the model outputs and supplemented using information from Table 9.A.1 of **?**.

atmosphere model. The NCAR CCSM4 model has been superseded by the CESM1 model, and so was not included. The two NorESM1 models are also very closely related to the CESM1 model, so was excluded. The BNU-ESM and FIO-ESM models were also excluded since they use outdated and low resolution versions of the CAM atmosphere included in the CESM1 model. The two models submitted from the CMCC are both based on an old atmosphere component and a very old ocean component. They also lack a full land surface model, therefore neither model was included. The CNRM-CM5 model and EC-EARTH models are very closely related, but EC-EARTH model includes more RCP4.5 runs so was retained over CNRM-CM5. The models from NOAA-GFDL differ primarily in their ocean component. GFDL-ESM2G uses the GOLD ocean model, while GFDL-ESM2M and GFDL-CM3 use the MOM4.1 ocean model. However, the MOM4 ocean model is also used in the models from the BCC, so GFDL-ESM2M and GFDL-CM3 are excluded. The NASA GISS-E2-R model was retained over the GISS-E2-H for the increased number of levels in the ocean model. The MOHC model in its HadGEM2-CC configuration has a relatively low resolution ocean component compared to most of the other models, so it is excluded in favor of the HadGEM2-ES configuration. The model submitted by NIMR/KMA is another version of the MOHC model, and so was excluded. The resolution of the atmospheric component of the IPSL-CM5A-LR model is also low compared to the rest of the ensemble, so it is excluded in favor of the IPSL-CM5A-MR configuration. Similarly, the atmospheric resolution of the MIROC models in their MIROC-ESM configuration is relatively low, so they are excluded and MIROC5 is retained. In contrast, the MPI-ESM-MR configuration features a very high resolution ocean component compared to the rest of the ensemble. Therefore the MPI-ESM-LR configuration is retained instead.

The thinned ensemble contains a total of 89 runs from 13 models, 50 runs from the historical scenario and 39 from the RCP4.5 scenario. The models included in the thinned

Table 2: Models included in the thinned ensemble. Number of runs available from each model for the historical and future time periods.

Modeling center	Model	Runs	
		Historical	Future
		N_{Hm}	N_{Fm}
BCC	BCC-CSM1.1(m)	3	1
CCCMA	CanESM2	5	5
NSF-DOE-NCAR	CESM1(CAM5)	3	3
ICHEC	EC-EARTH	8	9
LASG-CESS	FGOALS-g2	5	1
NOAA GFDL	GFDL-ESM2G	1	1
NASA GISS	GISS-E2-R	6	6
MOHC	HadGEM2-ES	4	4
INM	INM-CM4	1	1
IPSL	IPSL-CM5A-MR	3	1
MIROC	MIROC5	5	3
MPI-M	MPI-ESM-LR	3	3
MRI	MRI-CGCM3	3	1
Total		50	39

ensemble and the number of runs from each is listed in Table 2.

3 Estimating observation uncertainty

Estimates of the observation uncertainty σ_Z^2 are often not readily available. Several modeling centers produce "reanalysis" products that combine multiple observation sources using complex data assimilation techniques and numerical weather models. Given multiple reanalysis data sets we can approximate our uncertainty about the observed state of the climate. Let W_i be the output of reanalysis i, which we model as

$$W_i \sim N\left(\mu_W, \sigma_W^2\right) \tag{1}$$

where μ_W is interpreted as a representative reanalysis and the variance σ_W^2 quantifies the spread of the reanalyses. We expect the representative reanalysis μ_W to be similar to the actualized climate Y_{Ha}, and so we model the representative reanalysis as

$$\mu_W \sim N\left(Y_{Ha}, \sigma_{\Delta_W}^2\right) \tag{2}$$

The variance $\sigma_{\Delta_W}^2$ quantifies our uncertainty about the discrepancy between the representative reanalysis and the actual climate, due to sparsity of observations, errors in the numerical weather models etc. Similar to the models, we judge that the representative reanalysis is less like the actualized climate than the individual reanalyses are like the representative reanalysis, so we set

$$\sigma_{\Delta_W}^2 = \kappa_W^2 \sigma_W^2 \qquad\qquad \kappa_W \geq 1. \tag{3}$$

Conditioning the representative reanalysis μ_W on the actualized climate Y_{Ha} in Equation 2 induces a correlation (dependence) between the models and the reanalyses, i.e, $\text{cov}\left(W_i, X_{Hm}\right) = \text{var}\left(\mu_H\right) + \sigma_{\Delta_H}^2 + \sigma_a^2 + \sigma_{\Delta_W}^2$ for all $\{i, m\}$. Such a correlation makes sense, since climate models and reanalyses are very closely related, sharing very similar numerical cores.

For the analysis of Arctic surface temperature we combined four contemporary reanalysis data sets: ERA-Interim (**?**); NCEP CFSR (**?**); JRA-25 (**?**); and NASA MERRA (**?**). The coefficient κ_W was set to equal to the ensemble coefficient *kappa* at 1.2.

4 Derivation of Gibbs'-Metropolis updating equations

For the purposes of computation it is more convenient to work with precisions than variances, so let

$$\tau_m = 1/\sigma_m^2 \text{ for } m = 1, \ldots, M; \quad \phi_m = 1/\varphi_m^2 \text{ for } m = 1, \ldots, M$$
$$\tau_H = 1/\sigma_H^2; \qquad \tau_{F|H} = 1/\sigma_{F|H}^2; \qquad \tau_a = 1/\sigma_a^2; \qquad \phi_a = 1/\varphi_a^2; \qquad \tau_W = 1/\sigma_W^2.$$

The complete model defined by Equations 1–8 of the main text and Equations 1–3 of the supplementary material can be rewritten as

$$\begin{aligned}
X_{Hmr} \mid X_{Hm} &\sim N\left(X_{Hm}, \tau_m^{-1}\right) & X_{Fmr} \mid X_{Fm} &\sim N\left(X_{Fm}, (\phi_m \tau_m)^{-1}\right) \\
X_{Hm} &\sim N\left(\mu_H, \tau_H^{-1}\right) & X_{Fm} \mid X_{Hm} &\sim N\left(\mu_F + \beta\left(X_{Hm} - \mu_H\right), \tau_{F|H}^{-1}\right) \\
\tau_m &\sim Gamma\left(\frac{\nu_H}{2}, \frac{\nu_H \psi^2}{2}\right) & \phi_m &\sim Gamma\left(\frac{\nu_F}{2}, \frac{\nu_F \theta^2}{2}\right) \\
Y_{Ha} \mid Y_H &\sim N\left(Y_H, \tau_a^{-1}\right) & Y_{Fa} \mid Y_F &\sim N\left(Y_F, (\phi_a \tau_a)^{-1}\right) \\
Y_H &\sim N\left(\mu_H, \tau_{\Delta_H}^{-1}\right) & Y_F \mid Y_H &\sim N\left(\mu_F + \beta\left(Y_H - \mu_H\right), \tau_{\Delta_{F|H}}^{-1}\right) \\
\tau_a &\sim Gamma\left(\frac{\nu_{Ha}}{2}, \frac{\nu_{Ha} \psi^2}{2}\right) & \phi_a &\sim Gamma\left(\frac{\nu_{Fa}}{2}, \frac{\nu_{Fa} \theta^2}{2}\right) \\
W_i &\sim N\left(\mu_W, \tau_W^{-1}\right) & \mu_W &\sim N\left(Y_{Ha}, \tau_{\Delta_W}^{-1}\right)
\end{aligned}$$

where

$$\tau_{\Delta_H} = \tau_H/\kappa^2; \quad \tau_{\Delta_{F|H}} = \tau_{F|H}/\kappa^2; \quad \tau_{\Delta_W} = \tau_W/\kappa_W^2; \quad \nu_{Ha} = \nu_H/\kappa^2; \quad \nu_{Fa} = \nu_F/\kappa^2.$$

Vague conjugate prior probability distributions were specified for the parameters as follows $\mu_H, \mu_W, \beta \sim N\left(0, 10^6\right)$, $\mu_F \mid \mu_H \sim N\left(\mu_H, 10^6\right)$, $\tau_H, \tau_{F|H}, \tau_W \sim Inv\text{-}gamma\left(10^{-3}, 10^{-3}\right)$, $\psi^2, \theta^2 \sim Gamma\left(10^{-3}, 10^{-3}\right)$, and $\nu_H, \nu_F \sim Exp\left(1/M\right)$. The resulting full conditional posterior distributions all have standard forms with the exception of the degrees-of-freedom ν_H

and ν_F. Therefore, posterior inference can be efficiently accomplished by Gibbs' sampling with Metropolis-Hastings steps for ν_H and ν_F.

Let $\mathbf{X} = (X_{tmr}, s \in \{H, F\}, m = 1, \ldots, M, r = 1, \ldots, R_{tm})'$ be the model outputs, $\mathbf{Y} = (Y_H, Y_{Ha}, \tau_a)'$ be the latent state of the climate system, $\boldsymbol{\theta} = (\mu_H, \mu_F, \beta, \tau_H, \tau_{F|H}, \psi^2, \phi^2, \nu_H, \nu_F)'$ be the ensemble parameters, $\boldsymbol{\chi} = (X_{Hm}, X_{Fm}, \tau_m, \phi_m, m = 1, \ldots, M)'$ be the latent model states, $\mathbf{W} = (W_i, i = 1, \ldots, N)$ be the reanalysis outputs, and $\boldsymbol{\omega} = (\mu_W, \tau_W)'$ be the reanalysis parameters. The future state of the climate system defined by Y_F, Y_{Fa} and ϕ_a are purely predictive quantities and can be sampled after sampling of all other quantities is complete, using the equations above.

The joint posterior can be decomposed as

$$\Pr(\mathbf{Y}, \boldsymbol{\chi}, \boldsymbol{\theta}, \boldsymbol{\omega} \mid \mathbf{X}, \mathbf{W}) \propto \Pr(\mathbf{W} \mid \boldsymbol{\omega}) \Pr(\mathbf{Y} \mid \boldsymbol{\theta}, \boldsymbol{\omega}) \Pr(\mathbf{X} \mid \boldsymbol{\chi}) \Pr(\boldsymbol{\chi} \mid \boldsymbol{\theta}) \Pr(\boldsymbol{\theta}) \Pr(\boldsymbol{\omega})$$

The likelihood of the reanalysis outputs $\mathbf{W}$ given the reanalysis parameters $\boldsymbol{\omega}$ is proportional to

$$\Pr(\mathbf{W} \mid \boldsymbol{\omega}) \propto \prod_{i=1}^{N} \tau_W^{1/2} \exp\left(-\frac{\tau_W}{2}(W_i - \mu_W)^2\right).$$

The likelihood of the system $\mathbf{Y}$ given the ensemble parameters $\boldsymbol{\theta}$ and the reanalysis parameters $\boldsymbol{\omega}$ is proportional to

$$\Pr(\mathbf{Y} \mid \boldsymbol{\theta}, \boldsymbol{\omega}) \propto \tau_{\Delta_H}^{1/2} \exp\left(-\frac{\tau_{\Delta_H}}{2}(Y_H - \mu_H)^2\right) \tau_a^{1/2} \exp\left(-\frac{\tau_a}{2}(Y_{Ha} - Y_H)^2\right)$$
$$\frac{\left(\frac{\nu_{Ha}\psi^2}{2}\right)^{\nu_{Ha}/2}}{\Gamma(\nu_{Ha}/2)} \tau_a^{\nu_{Ha}/2-1} \exp\left(-\frac{\nu_{Ha}\psi^2}{2}\tau_a\right).$$

The likelihood of the model outputs $\mathbf{X}$ given the latent model states $\boldsymbol{\chi}$ is proportional to

$$\begin{aligned}\Pr\left(\mathbf{X} \mid \boldsymbol{\chi}\right) \propto & \prod_{m=1}^{M} \prod_{r=1}^{R_{Hm}} \tau_m^{1/2} \exp\left(-\frac{\tau_m}{2}\left(X_{Hmr} - X_{Hm}\right)^2\right) \\ & \prod_{m=1}^{M} \prod_{r=1}^{R_{Fm}} \left(\phi_m \tau_m\right)^{1/2} \exp\left(-\frac{\phi_m \tau_m}{2}\left(X_{Fmr} - X_{Fm}\right)^2\right).\end{aligned}$$

The likelihood of the model states $\boldsymbol{\chi}$ given the ensemble parameters θ is proportional to

$$\begin{aligned}\Pr\left(\boldsymbol{\chi} \mid \boldsymbol{\theta}\right) \propto & \prod_{m=1}^{M} \tau_H^{1/2} \exp\left(-\frac{\tau_H}{2}\left(X_{Hm} - \mu_H\right)^2\right) \\ & \prod_{m=1}^{M} \tau_F^{1/2} \exp\left(-\frac{\tau_F}{2}\left(X_{Fm} - \mu_F - \beta\left(X_{Hm} - \mu_H\right)\right)^2\right) \\ & \prod_{m=1}^{M} \frac{\left(\frac{\nu_H \psi^2}{2}\right)^{\nu_H/2}}{\Gamma\left(\nu_H/2\right)} \tau_m^{\nu_H/2-1} \exp\left(-\frac{\nu_H \psi^2}{2} \tau_m\right) \\ & \prod_{m=1}^{M} \frac{\left(\frac{\nu_F \theta^2}{2}\right)^{\nu_F/2}}{\Gamma\left(\nu_F/2\right)} \phi_m^{\nu_F/2-1} \exp\left(-\frac{\nu_F \theta^2}{2} \phi_m\right).\end{aligned}$$

The joint prior distribution of the ensemble parameters $\boldsymbol{\theta}$ is proportional to

$$\begin{aligned}\Pr\left(\boldsymbol{\theta}\right) \propto & \exp\left(-\frac{b_{\mu_H}}{2}\left(\mu_H - a_{\mu_H}\right)^2\right) \exp\left(-\frac{b_{\mu_F}}{2}\left(\mu_F - \mu_H\right)^2\right) \exp\left(-\frac{b_\beta}{2}\left(\beta - a_\beta\right)^2\right) \\ & \tau_H^{a_{\tau_H}-1} \exp\left(-b_{\tau_H}\tau_H\right) \tau_F^{a_{\tau_F}-1} \exp\left(-b_{\tau_F}\tau_F\right) \nu_H^{a_{\nu_H}-1} \exp\left(-b_{\nu_H}\nu_H\right) \nu_F^{a_{\nu_F}-1} \exp\left(-b_{\nu_F}\nu_F\right) \\ & \left(\psi^2\right)^{a_{\psi^2}-1} \exp\left(-b_{\psi^2}\psi^2\right) \left(\theta^2\right)^{a_{\theta^2}-1} \exp\left(-b_{\theta^2}\theta^2\right).\end{aligned}$$

The joint prior distribution of the reanalysis parameters $\boldsymbol{\omega}$ is proportional to

$$\Pr\left(\boldsymbol{\omega}\right) \propto \tau_{\Delta_W}^{1/2} \exp\left(-\frac{\tau_{\Delta_W}}{2}\left(\mu_W - Y_{Ha}\right)^2\right) \tau_W^{a_{\tau_W}-1} \exp\left(-b_{\tau_W}\tau_W\right).$$

The full conditional distributions of the system quantities $\mathbf{Y}$ are

$$
\begin{aligned}
Y_{Ha} \mid \ldots &\sim N\left(\frac{\tau_a Y_H + \tau_{\Delta_W}\mu_W}{\tau_a + \tau_{\Delta_W}}, \left(\tau_a + \tau_{\Delta_W}\right)^{-1}\right) \\
Y_H \mid \ldots &\sim N\left(\frac{\tau_{\Delta_H}\mu_H + \tau_a Y_{Ha}}{\tau_{\Delta_H} + \tau_a}, \left(\tau_{\Delta_H} + \tau_a\right)^{-1}\right) \\
\tau_a \mid \ldots &\sim Gamma\left(\frac{\nu_{Ha}+1}{2}, \frac{\nu_{Ha}\psi^2 + \left(Y_{Ha} - Y_H\right)^2}{2}\right)
\end{aligned}
$$

The full conditional distributions of the reanalysis parameters $\boldsymbol{\omega}$ are

$$
\begin{aligned}
\mu_W \mid \ldots &\sim N\left(\frac{\tau_W \sum_i W_i + \tau_{\Delta_W} Y_{Ha}}{\tau_W N + \tau_{\Delta_W}}, \left(\tau_W N + \tau_{\Delta_W}\right)^{-1}\right) \\
\tau_W \mid \ldots &\sim Gamma\left(a_{\tau_W} + \frac{N+1}{2}, b_{\tau_W} + \frac{1}{2}\sum_i \left(W_i - \mu_W\right)^2 + \frac{1}{2}\kappa_W^{-2}\left(\mu_W - Y_{Ha}\right)^2\right)
\end{aligned}
$$

The full conditional distributions of the latent model states $\boldsymbol{\chi}$ are

$$
\begin{aligned}
X_{Fm} \mid \ldots &\sim N\left(\frac{\tau_F\left(\mu_F + \beta\left(X_{Hm} - \mu_H\right)\right) + \phi_m\tau_m\sum_r X_{Fmr}}{\tau_F + \phi_m\tau_m R_{Fm}}, \left(\tau_F + \phi_m\tau_m R_{Fm}\right)^{-1}\right) \\
X_{Hm} \mid \ldots &\sim N\left(\frac{\tau_H\mu_H + \tau_F\beta\left(X_{Fm} - \mu_F + \beta\mu_H\right) + \tau_m\sum_r X_{Hmr}}{\tau_H + \tau_F\beta^2 + \tau_m R_{Hm}}, \left(\tau_H + \tau_F\beta^2 + \tau_m R_{Hm}\right)^{-1}\right) \\
\tau_m \mid \ldots &\sim Gamma\left(\frac{\nu_H + N_{Hm} + N_{Fm}}{2}, \frac{\nu_H\psi^2 + \sum_r\left(X_{Hmr} - X_{Hm}\right)^2 + \phi_m\sum_r\left(X_{Fmr} - X_{Fm}\right)^2}{2}\right) \\
\phi_m \mid \ldots &\sim Gamma\left(\frac{\nu_F + N_{Fm}}{2}, \frac{\nu_F\theta^2 + \tau_m\sum_r\left(X_{Fmr} - X_{Fm}\right)^2}{2}\right)
\end{aligned}
$$

The full conditional distributions of the ensemble parameters $\boldsymbol{\theta}$ are

$$\mu_H \mid \ldots \sim N\left(\tilde{\mu}_H, \left(b_{\mu_H} + b_{\mu_F} + \tau_H M + \tau_F \beta^2 M + \tau_{\Delta_H}\right)^{-1}\right)$$

$$\mu_F \mid \ldots \sim N\left(\frac{b_{\mu_F}\mu_H + \tau_F \sum_m \left(X_{Fm} - \beta\left(X_{Hm} - \mu_H\right)\right)}{b_{\mu_F} + \tau_F M}, \left(b_{\mu_F} + \tau_F M\right)^{-1}\right)$$

$$\beta \mid \ldots \sim N\left(\frac{b_\beta a_\beta + +\tau_F \sum_m \left(X_{Hm} - \mu_H\right)\left(X_{Fm} - \mu_F\right)}{b_\beta + \tau_F \sum_m \left(X_{Hm} - \mu_H\right)^2}, \left(b_\beta + \tau_F \sum_m \left(X_{Hm} - \mu_H\right)^2\right)^{-1}\right)$$

$$\tau_H \mid \ldots \sim Gamma\left(a_{\tau_H} + \frac{M+1}{2}, b_{\tau_H} + \frac{\sum_m \left(X_{Hm} - \mu_H\right)^2 + \kappa^{-2}\left(Y_H - \mu_H\right)^2}{2}\right)$$

$$\tau_F \mid \ldots \sim Gamma\left(a_{\tau_F} + \frac{M}{2}, b_{\tau_F} + \frac{\sum_m \left(X_{Fm} - \mu_F - \beta\left(X_{Hm} - \mu_H\right)\right)^2}{2}\right)$$

$$\psi^2 \mid \ldots \sim Gamma\left(a_{\psi^2} + \frac{\nu_H M + \nu_{Ha}}{2}, b_{\psi^2} + \frac{\nu_H \sum_m \tau_m + \nu_{Ha}\tau_a}{2}\right)$$

$$\theta^2 \mid \ldots \sim Gamma\left(a_{\theta^2} + \frac{\nu_F M}{2}, b_{\theta^2} + \frac{\nu_F \sum_m \phi_m}{2}\right)$$

where

$$\tilde{\mu}_H = \frac{b_{\mu_H} a_{\mu_H} + b_{\mu_F}\mu_F + \tau_H \sum_m X_{Hm} - \tau_F \beta \sum_m \left(X_{Fm} - \mu_F - \beta X_{Hm}\right) + \tau_{\Delta_H} Y_H}{b_{\mu_H} + b_{\mu_F} + \tau_H M + \tau_F \beta^2 M + \tau_{\Delta_H}}.$$

The full conditional distributions of the degrees-of-freedom ν_H and ν_F do not correspond to any standard distribution. The likelihoods associated with ν_H and ν_F are

$$l(\nu_H) = \frac{\beta_{Ha}^{\alpha_{Ha}}}{\Gamma(\alpha_{Ha})}\tau_a^{\alpha_{Ha}-1}\exp(-\beta_{Ha}\tau_a)\prod_m \frac{\beta_H^{\alpha_H}}{\Gamma(\alpha_H)}\tau_m^{\alpha_H-1}\exp(-\beta_H\tau_m)$$

and

$$l(\nu_F) = \prod_m \frac{\beta_F^{\alpha_F}}{\Gamma(\alpha_F)}\phi_m^{\alpha_F-1}\exp(-\beta_F\phi_m)$$

where

$$\alpha_{Ha} = \nu_{Ha}/2, \quad \beta_{Ha} = \nu_{Ha}\psi^2/2, \quad \alpha_H = \nu_H/2, \quad \beta_H = \nu_H\psi^2/2, \quad \alpha_F = \nu_F/2, \quad \beta_F = \nu_F\theta^2/2.$$

The prior densities of ν_H and ν_F are

$$p(\nu_H) \propto \nu_H^{a_{\nu_H}-1}\exp(-b_{\nu_H}\nu_H) \quad \text{and} \quad p(\nu_F) \propto \nu_F^{a_{\nu_F}-1}\exp(-b_{\nu_F}\nu_F).$$

The posterior distributions of ν_H and ν_F conditional on the current state of the other parameters can be sampled using the Metropolis-Hastings algorithm. For each $s \in \{H, F\}$:

1. Sample a new state $\nu_t^\star$ from $q(\nu_t^\star \mid \nu_t)$;

2. Calculate the Hastings ratio

$$r(\nu_t^\star, \nu_t) = \frac{l(\nu_t^\star)p(\nu_t^\star)q(\nu_t^\star \mid \nu_t)}{l(\nu_t)p(\nu_t)q(\nu_t \mid \nu_t^\star)};$$

3. Accept the new state $\nu_t^\star$ with probability

$$a(\nu_t^\star, \nu_t) = \min(1, r(\nu_t^\star, \nu_t)).$$

where $q(\nu_t^\star \mid \nu_t) = Gamma(\nu_t\lambda_t, \lambda_t)$ is the proposal distribution, with expectation ν_t and variance controlled by the free parameter λ_t. The acceptance rate of the Metropolis step can be controlled using the parameter λ_t.

5 Posterior computation

Four parallel chains were initialized for each grid box, from over-dispersed starting points. Initially, 20 000 samples were performed by each chain for each grid box. The first 10 000 samples were discarded as burn-in, and Gelman-Rubin diagnostics performed on the remaining 10 000 samples (**?**). If any random quantity had a potential scale reduction factor greater than 1.10, then sampling was continued for a further 10 000 samples per chain and diagnostics performed again until satisfactory convergence was indicated. We store every 40th sample from the last 10 000 samples of each chain, leading to a final sample size of 1000 for each grid box. The Metropolis-within-Gibbs' sampler was implemented in the R statistical computing language (**?**). Computation time for four parallel chains of 20 000 samples at a single grid box is around 5.5 s on a standard Linux workstation. The samplers for all grid boxes converged successfully. Convergence was achieved after the initial 20 000 samples at 50 % of grid boxes. Less than 2 % of grid boxes required more than 100 000 samples before convergence.

Inspection of the posterior distributions showed that, despite the small ensemble size, the ensemble parameters μ_H, μ_F, β, σ_H^2, $\sigma_{F|H}^2$ and ψ^2 and θ^2 are all very well constrained by the data. As expected, the degrees-of-freedom ν_H and ν_F were only mildly constrained compared to the exponential prior. The inter-quartile range (IQR) for the mode of ν_H over the 2880 grid boxes was 5–10, and for ν_F the IQR was 5–8, compared to the mode of zero for the exponential prior. However, both ν_H and ν_F tended to have long tails at individual grid boxes. Due to the extremely small sample size, the reanalysis spread σ_W was relatively poorly constrained compared to the other parameters, but the posterior mean was below 2.0 °C at more than 75 % of grid boxes.

Monte Carlo standard errors were computed for each parameter at each grid box (**??**).

The Monte Carlo standard error rarely accounted for more than 4.3 % of the posterior standard error, or exceeded 3.8 % of the absolute posterior mean.

Examination of correlation matrices for the posterior samples revealed that only the means μ_H and μ_F are consistently highly correlated (IQR Cor (μ_H, μ_F) 0.69–0.93), which is to be expected given the relationship in Equation 2 of the main text. Unsurprisingly, the internal variability ψ^2 and θ^2 are also moderately correlated (IQR -0.44–-0.37). The only other parameters to have consistently non-zero correlation in the posterior samples were ψ^2 and ν_H (IQR 0.13–0.28), and θ^2 and ν_F (IQR 0.08–0.16). Again, this is not surprising given the close relationship between these parameters in Equation 3 of the main text, and the small number of initial condition runs available from each model. None of these findings is particularly troubling, and so we conclude that the posterior simulation worked well.

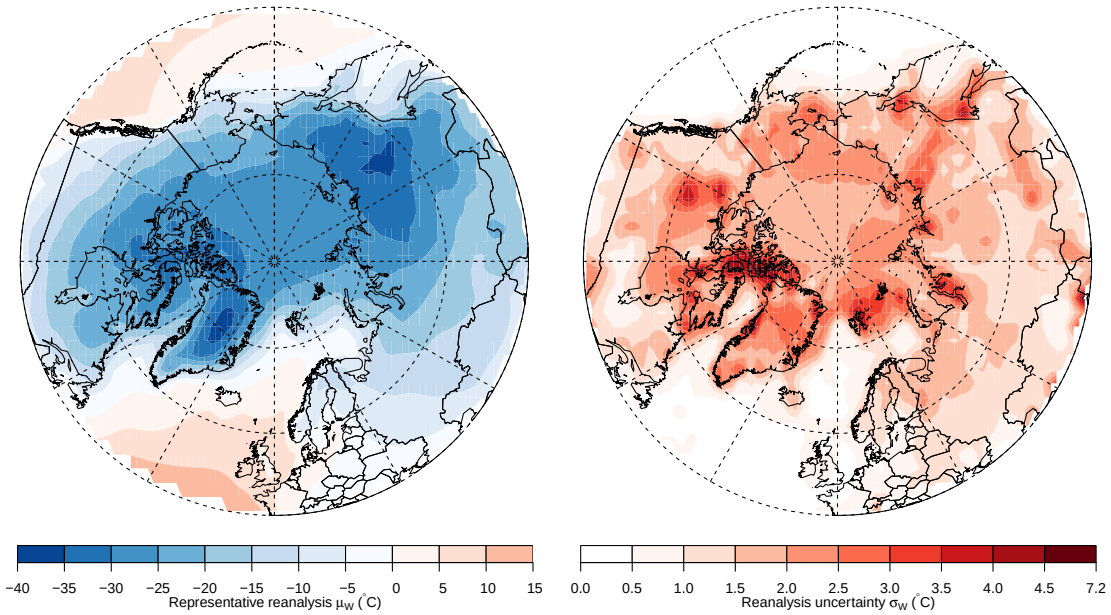

Figure 1: Posterior mean estimates of the representative reanalysis μ_W and the reanalysis uncertainty σ_W.

6 Posterior parameter estimates

The spread between the reanalyses σ_W is greater over land than over the ocean where temperatures vary more slowly (Figure 1). The reanalysis uncertainty increases with latitude as the number of observing stations decreases and the terrain tends to become more mountainous (Figure 1). The spread between the reanalyses is particularly large around the sea ice edge.

The representative historical climate μ_H is quite similar to the representative reanalysis μ_W, except over the Arctic ocean where climate models tend to be cold biased (Figure 2). The historical spread between the models σ_H is generally greater than the spread between the reanalyses σ_W (Figure 2). Like the reanalyses, the model spread tends to be greatest

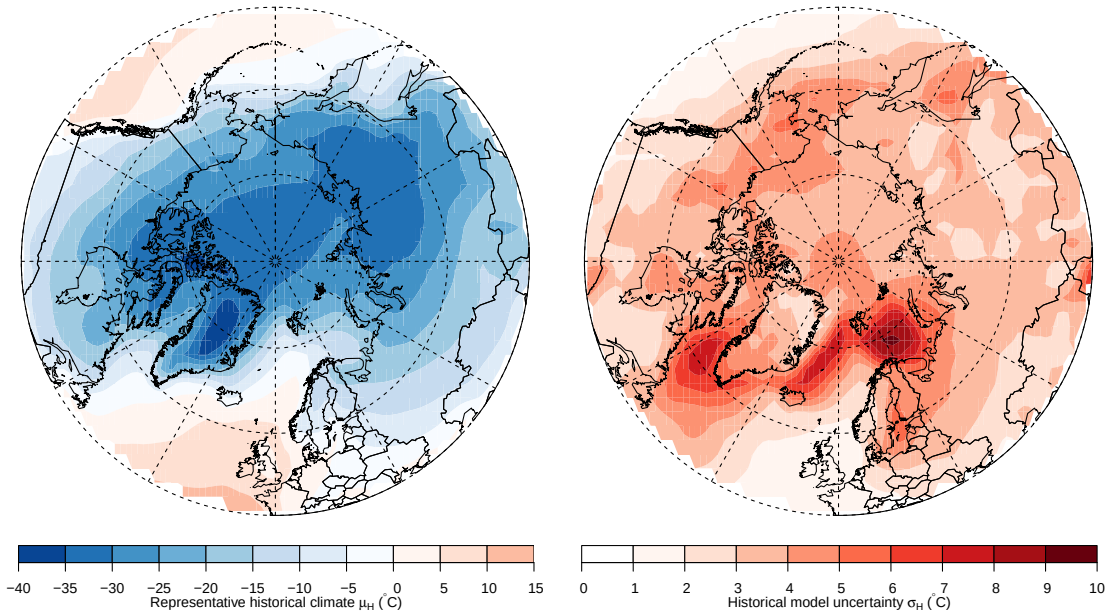

Figure 2: Posterior mean estimates of the representative historical climate μ_H and the historical model uncertainty σ_H.

over mountainous regions and near the sea ice edge.

The model response uncertainty $\sigma_{F|H}$ is greatest over the Arctic ocean, particularly to the east of Svalbard (Figure 3).

Like the reanalysis uncertainty, the representative internal variability ψ is greater over land than over the oceans, and highest in mountainous regions and close to the sea ice edge (Figure 4). The representative change in internal variability θ is small over most of the study area (Figure 4). Internal variability decreases close to the historical sea ice edge, where rising temperatures cause the ice edge to retreat and temperatures to stabilize. The climate in the interior of the Arctic becomes more variable as rising temperatures causes seasonal melting in regions permanently covered by sea ice during the historical period.

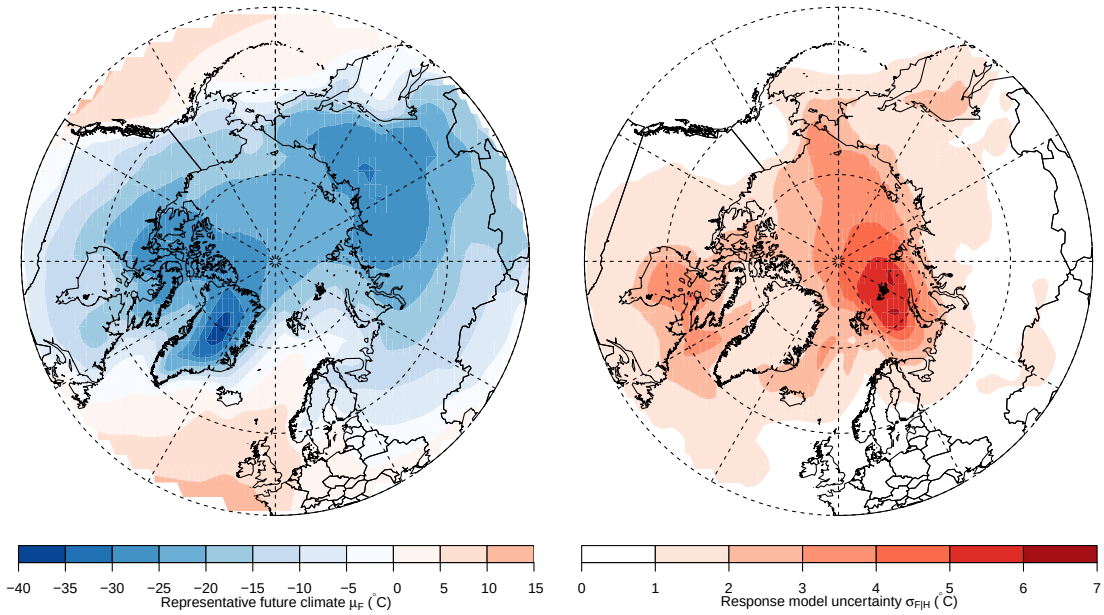

Figure 3: Posterior mean estimates of the representative future climate μ_F and the model response uncertainty $\sigma_{F|H}$

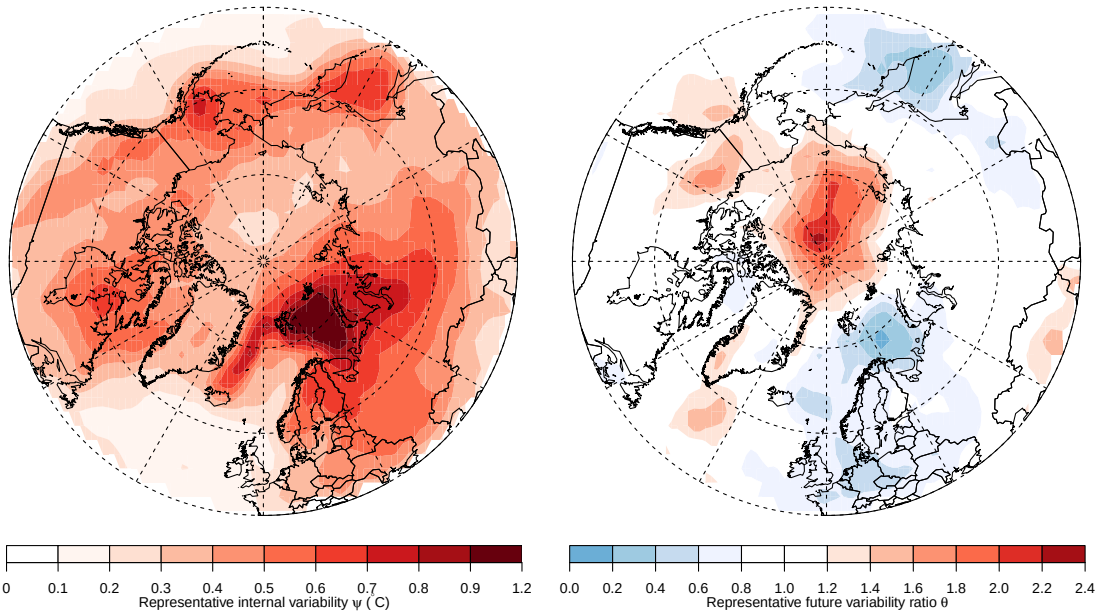

Figure 4: Posterior mean estimates of the representative historical internal variability ψ and change in internal variability θ.

References

Allen, M. R. and Ingram, W. J. (2002), 'Constraints on future changes in climate and the hydrologic cycle', *Nature* **419**(6903), 224–232.

Annan, J. D. and Hargreaves, J. C. (2010), 'Reliability of the CMIP3 ensemble', *Geophysical Research Letters* **37**, L02703.

Annan, J. D. and Hargreaves, J. C. (2011), 'Understanding the CMIP3 multimodel ensemble', *Journal of Climate* **24**(16), 4529–4538.

Bhat, K. S., Haran, M., Terando, A. and Keller, K. (2011), 'Climate Projections Using Bayesian Model Averaging and SpaceTime Dependence', *Journal of Agricultural, Biological, and Environmental Statistics* **16**(4), 606–628.

Bishop, C. H. and Abramowitz, G. (2013), 'Climate model dependence and the replicate Earth paradigm', *Climate Dynamics* **41**(3-4), 885–900.

Bowman, K. W., Cressie, N., Qu, X. and Hall, A. D. (2018), 'A hierarchical statistical framework for emergent constraints: application to snow-albedo feedback', *Geophysical Research Letters* **45**(23), 13050–13059.

Bracegirdle, T. J. and Stephenson, D. B. (2012), 'Higher precision estimates of regional polar warming by ensemble regression of climate model projections', *Climate Dynamics* **39**(12), 2805–2821.

Bracegirdle, T. J. and Stephenson, D. B. (2013), 'On the robustness of emergent constraints used in multimodel climate change projections of arctic warming', *Journal of Climate* **26**(2), 669–678.

Brient, F. (2020), 'Reducing uncertainties in climate projections with emergent constraints: Concepts, examples and prospects', *Advances in Atmospheric Sciences* **37**, 1–15.

Burke, E. J., Jones, C. D. and Koven, C. D. (2013), 'Estimating the permafrost-carbon climate response in the CMIP5 climate models using a simplified approach', *Journal of Climate* **26**(14), 4897–4909.

Buser, C. M., Künsch, H. R., Lüthi, D., Wild, M. and Schär, C. (2009), 'Bayesian multi-model projection of climate: Bias assumptions and interannual variability', *Climate Dynamics* **33**(6), 849–868.

Chandler, R. E. (2013), 'Exploiting strength, discounting weakness: combining information from multiple climate simulators', *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **371**, 20120388.

Collins, M. (2007), 'Ensembles and probabilities: A new era in the prediction of climate change', *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **365**(1857), 1957–1970.

Collins, M., Chandler, R. E., Cox, P. M., Huthnance, J. M., Rougier, J. and Stephenson, D. B. (2012), 'Quantifying future climate change', *Nature Climate Change* **2**(6), 403–409.

Cox, P. M., Huntingford, C. and Williamson, M. S. (2018), 'Emergent constraint on equilibrium climate sensitivity from global temperature variability', *Nature* **553**(7688), 319–322.

Craig, P. S., Goldstein, M., Rougier, J. C. and Seheult, A. H. (2001), 'Bayesian forecasting for complex systems using computer simulations', *Journal of the American Statistical Association* **96**(454), 717–729.

Cubasch, U., Meehl, G. A., Boer, G. J., Stouffer, R. J., Dix, M., Noda, A., Senior, C. A., Raper, S. and Yap, K. S. (2001), Projections of Future Climate Change, *in* J. Houghton, Y. Ding, D. J. Griggs, M. Noguer, P. J. van der Linden, K. Dai, X. and Maskell and C. A. Johnson, eds, 'Climate Change 2001: The Scientific Bases. Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change', Cambridge University Press, p. 881.

Deser, C., Phillips, A. S., Bourdette, V. and Teng, H. (2012), 'Uncertainty in climate change projections: the role of internal variability', *Climate Dynamics* **38**, 527–546.

Frost, C. and Thompson, S. G. (2000), 'Correcting for regression dilution bias: comparison of methods for a single predictor variable', *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **163**(2), 173–189.

Furrer, R., Sain, S. R., Nychka, D. W. and Meehl, G. A. (2007), 'Multivariate Bayesian analysis of atmosphere-ocean general circulation models', *Environmental and Ecological Statistics* **14**(3), 249–266.

Greene, A. M., Goddard, L. and Lall, U. (2006), 'Probabilistic multimodel regional temperature change projections', *Journal of Climate* **19**(17), 4326–4343.

Hall, A., Cox, P., Huntingford, C. and Klein, S. (2019), 'Progressing emergent constraints on future climate change', *Nature Climate Change* **9**(4), 269–278.

Hall, A. D. and Qu, X. (2006), 'Using the current seasonal cycle to constrain snow albedo feedback in future climate change', *Geophysical Research Letters* **33**(3), 1–4.

Hawkins, E. and Sutton, R. T. (2009), 'The Potential to Narrow Uncertainty in Regional Climate Predictions', *Bulletin of the American Meteorological Society* **90**, 1095–1107.

Hawkins, E. and Sutton, R. T. (2011), 'The potential to narrow uncertainty in projections of regional precipitation change', *Climate Dynamics* **37**(1-2), 407–418.

Holland, M. M. and Bitz, C. M. (2003), 'Polar amplification of climate change in coupled models', *Climate Dynamics* **21**(3-4), 221–232.

Jun, M., Knutti, R. and Nychka, D. W. (2008), 'Spatial Analysis to Quantify Numerical Model Bias and Dependence', *Journal of the American Statistical Association* **103**(483), 934–947.

Kennedy, M. C. and O'Hagan, A. (2001), 'Bayesian Calibration of Computer Models', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **63**(3), 425–464.

Knutti, R., Furrer, R., Tebaldi, C., Cermak, J. and Meehl, G. A. (2010), 'Challenges in combining projections from multiple climate models', *Journal of Climate* **23**(10), 2739–2758.

Knutti, R., Masson, D. and Gettelman, A. (2013), 'Climate model genealogy: Generation CMIP5 and how we got there', *Geophysical Research Letters* **40**(6), 1194–1199.

Knutti, R., Sedláček, J., Sanderson, B. M., Lorenz, R., Fischer, E. M. and Eyring, V. (2017), 'A climate model projection weighting scheme accounting for performance and interdependence', *Geophysical Research Letters* **44**(4), 1909–1918.

Koven, C. D., Riley, W. J. and Stern, A. (2013), 'Analysis of Permafrost Thermal Dynamics and Response to Climate Change in the CMIP5 Earth System Models', *Journal of Climate* **26**(6), 1877–1900.

Lambert, S. J. and Boer, G. J. (2001), 'CMIP1 evaluation and intercomparison of coupled climate models', *Climate Dynamics* **17**(2-3), 83–106.

Mahlstein, I. and Knutti, R. (2011), 'Ocean heat transport as a cause for model uncertainty in projected arctic warming', *Journal of Climate* **24**(5), 1451–1460.

Masson, D. and Knutti, R. (2011), 'Climate model genealogy', *Geophysical Research Letters* **38**(8), L08703.

McKinnon, K. A. and Deser, C. (2018), 'Internal Variability and Regional Climate Trends in an Observational Large Ensemble', *Journal of Climate* **31**(17), 6783–6802.

Min, S. K. and Hense, A. (2006), 'A Bayesian approach to climate model evaluation and multi-model averaging with an application to global mean surface temperatures from IPCC AR4 coupled climate models', *Geophysical Research Letters* **33**(8), L08708.

Moss, R. H., Edmonds, J. A., Hibbard, K. A., Manning, M. R., Rose, S. K., van Vuuren, D. P., Carter, T. R., Emori, S., Kainuma, M., Kram, T., Meehl, G. A., Mitchell, J. F. B., Nakicenovic, N., Riahi, K., Smith, S. J., Stouffer, R. J., Thomson, A. M., Weyant, J. P. and Wilbanks, T. J. (2010), 'The next generation of scenarios for climate change research and assessment', *Nature* **463**(7282), 747–756.

Northrop, P. J. and Chandler, R. E. (2014), 'Quantifying Sources of Uncertainty in Projections of Future Climate', *Journal of Climate* **27**(23), 8793–8808.

Oreskes, N., Shrader-Frechette, K. and Belitz, K. (1994), 'Verification, Validation, and Confirmation of Numerical Models in the Earth Sciences', *Science* **263**(5147), 641–646.

Parker, W. S. (2006), 'Understanding pluralism in climate modeling', *Foundations of Science* **11**(4), 349–368.

Pennell, C. and Reichler, T. (2011), 'On the Effective Number of Climate Models', *Journal of Climate* **24**(9), 2358–2367.

Poppick, A., McInerney, D. J., Moyer, E. J. and Stein, M. L. (2016), 'Temperatures in transient climates: Improved methods for simulations with evolving temporal covariances', *The Annals of Applied Statistics* **10**(1), 477–505.

Qu, X. and Hall, A. D. (2014), 'On the persistent spread in snow-albedo feedback', *Climate Dynamics* **42**(1-2), 69–81.

Räisänen, J. and Palmer, T. N. (2001), 'A probability and decision-model analysis of a multimodel ensemble of climate change simulations', *Journal of Climate* **14**(15), 3212–3226.

Rougier, J. C., Goldstein, M. and House, L. (2013), 'Second-Order Exchangeability Analysis for Multimodel Ensembles', *Journal of the American Statistical Association* **108**(503), 852–863.

Sanderson, B. M., Knutti, R. and Caldwell, P. M. (2015*a*), 'A representative democracy to reduce interdependency in a multimodel ensemble', *Journal of Climate* **28**(13), 5171–5194.

Sanderson, B. M., Knutti, R. and Caldwell, P. M. (2015*b*), 'Addressing interdependency in a multimodel ensemble by interpolation of model properties', *Journal of Climate* **28**(13), 5150–5170.

Shiogama, H., Emori, S., Hanasaki, N., Abe, M., Masutomi, Y., Takahashi, K. and Nozawa, T. (2011), 'Observational constraints indicate risk of drying in the Amazon basin.', *Nature communications* **2**, 253.

Slater, A. G. and Lawrence, D. M. (2013), 'Diagnosing present and future permafrost from climate models', *Journal of Climate* **26**(15), 5608–5623.

Smith, R. L., Tebaldi, C., Nychka, D. W. and Mearns, L. O. (2009), 'Bayesian Modeling of Uncertainty in Ensembles of Climate Models', *Journal of the American Statistical Association* **104**(485), 97–116.

Stainforth, D. A., Allen, M. R., Tredger, E. R. and Smith, L. A. (2007), 'Confidence, uncertainty and decision-support relevance in climate predictions', *Philosophical Transactions of the Royal Society A* **365**, 2145–2161.

Stephenson, D. B., Collins, M., Rougier, J. C. and Chandler, R. E. (2012), 'Statistical problems in the probabilistic prediction of climate change', *Environmetrics* **23**(5), 364–372.

Taylor, K. E., Stouffer, R. J. and Meehl, G. A. (2012), 'An overview of CMIP5 and the experiment design', *Bulletin of the American Meteorological Society* **93**(4), 485–498.

Tebaldi, C. and Knutti, R. (2007), 'The use of the multi-model ensemble in probabilistic climate projections', *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **365**(1857), 2053–2075.

Tebaldi, C. and Sansó, B. (2009), 'Joint projections of temperature and precipitation change from multiple climate models: A hierarchical Bayesian approach', *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **172**(1), 83–106.

Tebaldi, C., Smith, R. L., Nychka, D. W. and Mearns, L. O. (2005), 'Quantifying uncertainty in projections of regional climate change: A Bayesian approach to the analysis of multimodel ensembles', *Journal of Climate* **18**(10), 1524–1540.

Thompson, D. W. J., Barnes, E. A., Deser, C., Foust, W. E. and Phillips, A. S. (2015), 'Quantifying the role of internal climate variability in future climate trends', *Journal of Climate* **28**(16), 6443–6456.

Watterson, I. G. and Whetton, P. H. (2011), 'Distributions of decadal means of temperature and precipitation change under global warming', *Journal of Geophysical Research: Atmospheres* **116**(7), 1–13.

Weigel, A. P., Knutti, R., Liniger, M. A. and Appenzeller, C. (2010), 'Risks of model weighting in multimodel climate projections', *Journal of Climate* **23**(15), 4175–4191.

Yip, S., Ferro, C. A. T. and Stephenson, D. B. (2011), 'A Simple, Coherent Framework for Partitioning Uncertainty in Climate Predictions', *Journal of Climate* **24**(17), 4634–4643.