




## On Constraining Projections of Future Climate Using Observations and Simulations From Multiple Climate Models

Philip G. Sansom , David B. Stephenson & Thomas J. Bracegirdle


To cite this article: Philip G. Sansom , David B. Stephenson & Thomas J. Bracegirdle (2021): On Constraining Projections of Future Climate Using Observations and Simulations From Multiple Climate Models, Journal of the American Statistical Association, DOI: [10.1080/01621459.2020.1851696](https://doi.org/10.1080/01621459.2020.1851696)

To link to this article: <https://doi.org/10.1080/01621459.2020.1851696>

 View supplementary material [↗](#)

 Published online: 19 Jan 2021.

 Submit your article to this journal [↗](#)

 Article views: 120

 View related articles [↗](#)

 View Crossmark data [↗](#)



# On Constraining Projections of Future Climate Using Observations and Simulations From Multiple Climate Models

Philip G. Sansom<sup>a</sup>, David B. Stephenson<sup>a</sup>, and Thomas J. Bracegirdle<sup>b</sup>

<sup>a</sup>College of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter, UK; <sup>b</sup>British Antarctic Survey, Cambridge, UK

## ABSTRACT

Numerical climate models are used to project future climate change due to both anthropogenic and natural causes. Differences between projections from different climate models are a major source of uncertainty about future climate. Emergent relationships shared by multiple climate models have the potential to constrain our uncertainty when combined with historical observations. We combine projections from 13 climate models with observational data to quantify the impact of emergent relationships on projections of future warming in the Arctic at the end of the 21st century. We propose a hierarchical Bayesian framework based on a coexchangeable representation of the relationship between climate models and the Earth system. We show how emergent constraints fit into the coexchangeable representation, and extend it to account for internal variability simulated by the models and natural variability in the Earth system. Our analysis shows that projected warming in some regions of the Arctic may be more than 2 °C lower and our uncertainty reduced by up to 30% when constrained by historical observations. A detailed theoretical comparison with existing multi-model projection frameworks is also provided. In particular, we show that projections may be biased if we do not account for internal variability in climate model predictions. Supplementary materials for this article, including a standardized description of the materials available for reproducing the work, are available as an online supplement.

## ARTICLE HISTORY

Received November 2017  
Accepted November 2020

## KEYWORDS

Bayesian modeling; Coupled Model Intercomparison Project Phase 5; Emergent constraints; Hierarchical models; Measurement error

## 1. Introduction

Scientific inquiry into complex systems such as the climate naturally leads to multiple models of a system. The situation in climate science is unusual since different climate models are not treated as incompatible or competing. Instead, each model is treated as a plausible representation of the climate system (Parker 2006). This has led to the use of multi-model ensembles to quantify the uncertainty in projections of future climate introduced by choices in model design, usually referred to simply as model uncertainty (Tebaldi and Knutti 2007). Statistical methods are required to interpret projections from multi-model ensembles and to make credible probabilistic inferences about future climate change.

In addition to model uncertainty, projections of future climate are subject to a number of other sources of uncertainty. Model inadequacy refers to differences between the models and the Earth system, for example, missing processes (Craig et al. 2001; Stainforth et al. 2007). We intuitively think of climate as the distribution of weather (Stainforth et al. 2007; Stephenson et al. 2012). Natural variability refers to the range of possible conditions we might experience and is sometimes referred to as sampling uncertainty, since we only observe a single actualization of the Earth system (Chandler 2013). Climate models attempt to simulate natural variability by performing multiple simulations from slightly different initial conditions. This is known as internal variability or initial condition uncertainty. Climate projections are also subject to forcing uncertainty and

parameter uncertainty. Forcing uncertainty arises due to uncertainty about future emissions of greenhouse gases, both anthropogenic and natural. Parameter uncertainty refers to uncertainty about choice of the internal parameters in climate models (Collins 2007). Forcing uncertainty is usually circumvented by making projections rather than predictions of future climate, that is, predictions conditioned on an assumed future emissions scenario (e.g., Moss et al. 2010). The computational cost of running sufficiently large perturbed-parameter experiments to span the full range of parameter uncertainty for a single climate model can be prohibitive. Therefore, multi-model ensembles usually consist of a set of “best estimates,” that is, a single version of each model with the internal parameters fixed (Knutti et al. 2010).

In this article, we develop a hierarchical Bayesian framework for combining projections from multiple models, applied to projecting climate change in the Arctic at the end of the 21st century. The proposed framework separates model uncertainty and model inadequacy, and accounts for internal variability and natural variability in future projections. In addition, we are able to constrain projections of future climate using historical observations (where suitable constraints have been identified) while accounting for uncertainty in the observations.

To make projections of future climate from multi-model ensembles, it is necessary to make assumptions about the relationship between climate models and the Earth system. One

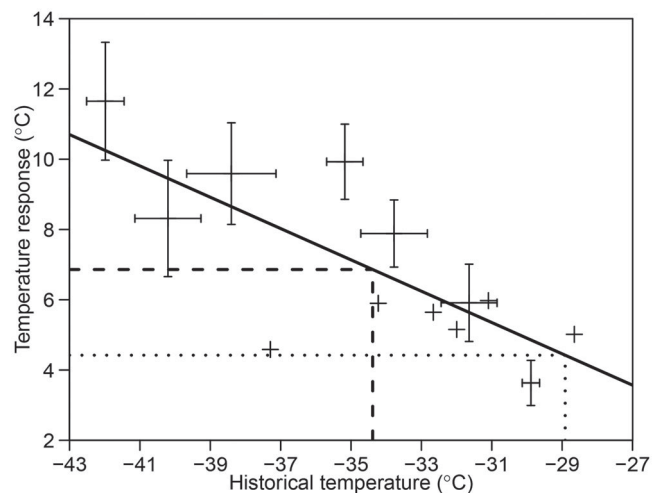
widely used assumption is that skill in reproducing past climate implies skill in projecting future climate. Climate scientists have long recognized that no single model will perform best for all variables or in all regions (Lambert and Boer 2001; Jun, Knutti, and Nychka 2008). Various approaches have been proposed for weighting projections from multiple climate models based on their ability to reproduce past climate, these include heuristics (Sanderson, Knutti, and Caldwell 2015a, 2015b, Knutti et al. 2017), multiple regression (Greene, Goddard, and Lall 2006; Bishop and Abramowitz 2013), pattern scaling (Shiogama et al. 2011; Watterson and Whetton 2011), and Bayesian model averaging (Min and Hense 2006; Bhat et al. 2011). However, Weigel et al. (2010) demonstrated that weights that do not accurately reflect the projection skill of the models can lead to less reliable projections than weighting all models equally. Long-term climate projection involves extrapolation to states that have not been observed in recent Earth history. Therefore, the ability to reproduce observed data does not guarantee skill for projecting future events (Oreskes, Shrader-Frechette, and Belitz 1994). However, we should certainly be cautious when interpreting projections from models that are not able to adequately reproduce observed data, although how such performance should be quantified remains an open question (Knutti et al. 2010).

Weighting all models equally implies that each climate model performs equally well for simulating future climate *change*. This has led to the alternative assumption that any bias between the models and the Earth system remains approximately constant over time (Buser et al. 2009). Under this assumption, two main interpretations of multi-model experiments have emerged (Stephenson et al. 2012). The “truth plus error” approach treats the output of each model as the “true” state of the Earth system plus some error that is unique to each model (Cubasch et al. 2001; Tebaldi et al. 2005; Furrer et al. 2007; Smith et al. 2009; Tebaldi and Sansó 2009). The “exchangeable” approach treats the Earth system as though it were just another climate model, that is, our inferences about the future climate of the Earth system should be the same as for a climate model with an identical historical climate (Räisänen and Palmer 2001; Annan and Hargreaves 2010, 2011). Neither interpretation is entirely satisfactory. The truth-plus-error interpretation implies that we can improve the precision (but not necessarily the accuracy) of our projections of future climate simply by adding more models to our ensemble (Annan and Hargreaves 2010). The exchangeable interpretation ignores the inherent differences between computer models and the physical systems they seek to represent (Craig et al. 2001; Kennedy and O’Hagan 2001).

Both the truth-plus error and exchangeable approaches acknowledge differences between models, and between individual models and the Earth system. What is missing are differences from the Earth system that are *common* to all models. All climate models are based on a shared but limited knowledge of the Earth system and face similar technological constraints (e.g., similar numerical methods, available CPU time, memory, etc.), so common limitations will inevitably occur (Stainforth et al. 2007). To address this issue, Chandler (2013) and Rougier, Goldstein, and House (2013) independently introduced the idea of representing common model errors as a discrepancy between

the expected state of the Earth system and a “consensus” or “representative” model. This has the effect of separating model uncertainty (differences between models) from model inadequacy (common differences between models and the Earth system).

Historical observations have been used in a variety of ways to constrain projections from individual models (Collins et al. 2012). However, if systematic relationships existed between the historical states and climate responses simulated by *multiple* models, then it might be possible to constrain projections of future climate in a multi-model ensemble without assigning weights to individual models. One of the earliest examples of such a relationship was noted by Allen and Ingram (2002) who referred to it as an “emergent constraint,” since it emerged from analysis of a collection of model simulations rather than by direct calculation based on theory. There is now a growing body of evidence that such relationships may exist at the local or process level, and even at the global level (Hall et al. 2019). In general, we prefer the term “emergent relationship.” We reserve the term “emergent constraint” for when physical insight indicates that the relationship should also hold in the Earth system. Figure 1 shows an example of a well understood emergent constraint on surface temperature in the Arctic due to albedo feedbacks caused by variations in sea-ice coverage simulated by the models (Bracegirdle and Stephenson 2012). Other examples of emergent relationships have been found in the cryosphere, atmospheric chemistry, the carbon cycle and various other areas of the Earth system (Brient 2020). The constraint on Equilibrium Climate Sensitivity proposed by Cox, Huntingford, and Williamson (2018) is a rare example that was derived from theory, then found to be present in a collection of model simulations. Simple linear regression is often



**Figure 1.** Near-surface warming in the Canadian Arctic Archipelago. Thirty-year mean temperature change between 1975–2005 and 2069–2099 as simulated by an ensemble of 13 climate models under the RCP4.5 mid-range mitigation scenario, for a  $2.5^\circ \times 2.5^\circ$  grid box centered on Melville Island ( $76^\circ\text{N}, 111^\circ\text{W}$ ). Crosses mark the mean climate and climate response simulated by each model. Whiskers indicate the range of 30-year mean outcomes from the initial condition runs of each model. The dashed line indicates the mean climate and climate response of the ensemble. The solid line is a simple linear regression estimate of the emergent relationship between the climate response and the historical climate. The dotted line indicates the observed historical climate and projected climate response given the estimated emergent relationship.

used to estimate emergent relationships. However, projection either implicitly treats the Earth system as exchangeable with the models (e.g., Bracegirdle and Stephenson 2012, 2013), or simply excludes all models that fall outside the plausible range of the observations (e.g., Hall and Qu 2006; Qu and Hall 2014).

Multi-model ensembles are sometimes known as “ensembles of opportunity” since models are not systematically selected to span model uncertainty, and cannot be considered a random sample from some larger population (Stephenson et al. 2012). In particular, several research centers maintain more than one model, and models from different centers often share common components (Knutti, Masson, and Gettelman 2013). Similar models are likely to give similar outputs, leading to clustering that could result in biased inferences if not properly accounted for. This is especially important when analyzing emergent constraints since a large cluster of outlying models could strongly influence any regression relationship. Therefore, care is required to ensure that our assumptions in representing model uncertainty and inadequacy are satisfied.

Model uncertainty/inadequacy tends to dominate other sources of uncertainty in long-term climate projections (Hawkins and Sutton 2009; Yip, Ferro, and Stephenson 2011). However, there is now a significant body of work highlighting the importance of internal variability and natural variability (Deser et al. 2012; Thompson et al. 2015; McKinnon and Deser 2018). Several studies have shown that the contribution of internal variability is nonnegligible compared to model uncertainty for some variables at the global scale, and particularly at the regional scale (Hawkins and Sutton 2009, 2011; Northrop and Chandler 2014). The internal variability simulated by each model is indicated by the whiskers in Figure 1. Current frameworks for multi-model inference often ignore internal variability and select a single initial condition run from each model (e.g., Tebaldi et al. 2005; Smith et al. 2009; Bishop and Abramowitz 2013) or take the average over all runs from each model (e.g., Watterson and Whetton 2011; Bracegirdle and Stephenson 2012). Measurement and representation errors in our observations of the climate system can also contribute significant observation uncertainty. Some authors have accounted for observation uncertainty (e.g., Bowman et al. 2018; Cox, Huntingford, and Williamson 2018), but it is frequently ignored and plays an important role if we want to constrain future projections using past observations.

The remainder of this study proceeds as follows. Section 2 outlines the data used to project future warming in the Arctic. In Section 3, we develop a hierarchical Bayesian framework for inferring time mean future climate from multi-model experiments for any future time period and location for which we have model simulations, conditional on simulations of a recent period for which we have corresponding observations. We do not attempt to account for spatial correlation between locations or temporal variation within each time period. Section 4 compares our proposed framework to existing multi-model ensemble approaches. In Section 5, we apply our framework to the projection of future climate change in the Arctic. We end with concluding remarks in Section 6.

## 2. Future Climate Change in the Arctic

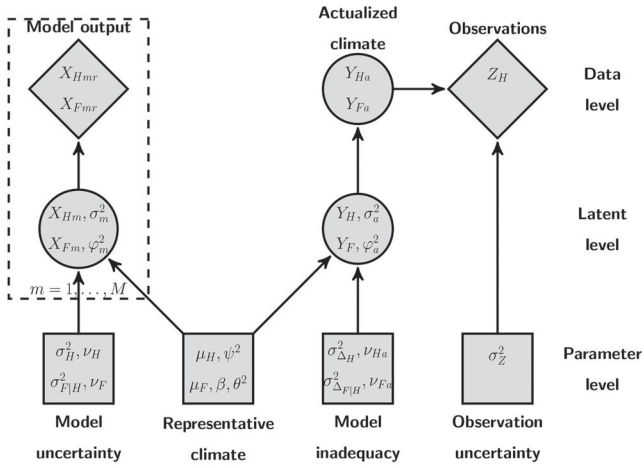
The magnitude of the projected warming in the polar regions is much greater than at lower latitudes (Holland and Bitz 2003). We combine outputs from 13 climate models participating in the World Climate Research Programme’s Coupled Model Inter-comparison Project Phase 5 (CMIP5, Taylor, Stouffer, and Meehl 2012) to investigate the impact of emergent constraints on projections of winter (December–January–February) near-surface (2 m) temperature change in the Arctic. The climate models included and the number of initial condition runs available from each are listed in the supplementary materials. We compare the 30 year average winter temperature between two time periods. The historical period is defined as between December 1975 and January 2005, as simulated under the CMIP5 historical emissions scenario. The future period of interest is between December 2069 and January 2099, as simulated under the RCP4.5 mid-range mitigation scenario (Moss et al. 2010). A total of 50 initial condition runs of the historical period were included, and 39 runs of the future period. The domain of interest is 45°N–90°N, including not only the Arctic Ocean but also the Bering Sea and the Sea of Okhotsk, both of which also currently experience significant seasonal sea ice coverage. Due to the presence of seasonal ice coverage and the complexity associated with modeling it, both model uncertainty and internal variability in near-surface temperature are much greater in the Arctic than at lower latitudes (Northrop and Chandler 2014). Prior to analysis, data from all models were interpolated bicubically to a common grid with equal 2.5° spacing in both longitude and latitude.

Observational data in the Arctic are very sparse and no spatially complete datasets exist that include estimates of observational uncertainty. Therefore, we combine four contemporary reanalysis datasets to obtain spatially complete estimates of surface temperature during the historical period, and estimates of the observational uncertainty (see the supplementary materials for details). Reanalysis data were interpolated to the same grid as the models.

## 3. A Hierarchical Framework for Multi-Model Experiments

The proposed framework is summarized in graphical form in Figure 2. We compare one historical time period denoted  $H$  and one future time period denoted  $F$ , conditioned on a single future emissions scenario. The top level of Figure 2 consists of quantities for which we have data, that is, model outputs ( $X_{Hmr}, X_{Fmr}$ ) and observations ( $Z_H$ ). The mid-level consists of the climates of the individual models and the Earth system, quantified by the means ( $X_{Hm}, X_{Fm}, Y_H, Y_F$ ) and variances ( $\sigma_m^2, \psi_m^2, \sigma_a^2, \psi_a^2$ ) of the simulated or plausible conditions during each time period. The bottom level consists of parameters quantifying model uncertainty, the “representative” climate of the ensemble, model inadequacy, and observation uncertainty. All of these quantities will be fully defined in the development that follows.

We proceed in three stages. First, we propose a hierarchical model for the outputs of the multi-model ensemble (left-hand side, Figure 2). Second, we propose a similar hierarchical model for the climate of the Earth system (middle right, Figure 2).



**Figure 2.** Graphical representation. The proposed framework represented as a directed acyclic graph. Diamonds represent data, circles represent latent quantities, and squares represent parameters. The dashed box represents the multi-model ensemble. The actualized climate  $Y_{ta}$  is placed at the data level to emphasize its relationship with the model runs  $X_{tmr}$ .

Finally, we specify a model for the relationship between the actualized climate and the observations (right-hand side, Figure 2).

### 3.1. The Multi-Model Ensemble

Suppose we have an ensemble of  $M$  climate models. Each model performs a number of runs of the historical and future time periods, conditioned on a single future emissions scenario. Each run is initialized from slightly perturbed initial conditions. Let  $X_{tmr}$  be the output of run  $r$ , during time period  $t = \{H, F\}$ , by model  $m = 1, \dots, M$ . The outputs  $X_{tmr}$  are assumed to be time averages over periods of equal length. The number of runs of each model for each time period is denoted  $R_{tm}$ , that is,  $r = 1, \dots, R_{tm}$  for model  $m$  in time period  $t$ . We do *not* require that the number of runs from each model be equal, or that the number of runs of each period by a particular model be equal (frequently  $R_{Fm} < R_{Hm}$ ). Each model is attempting to simulate the same target, that is, the climate of the Earth system under a specific emissions scenario. Therefore, we assume a priori that the model outputs  $X_{tmr}$  are exchangeable, conditional on the emissions scenario. Exchangeability implies that we hold the same prior beliefs about the output of every run from every model, given a particular scenario. Therefore, we should specify the same probability model for each run of a particular scenario from every model. We model the individual runs  $X_{tmr}$  as

$$X_{Hmr}|X_{Hm} \sim N(X_{Hm}, \sigma_m^2) \quad X_{Fmr}|X_{Fm} \sim N(X_{Fm}, (\varphi_m \sigma_m)^2). \quad (1)$$

The outputs  $X_{tmr}$  are assumed to be independent between runs  $r$ , conditional on the other parameters. The model specific means  $X_{tm}$  represent the expected climate of model  $m$  at time  $t$ . The model specific variances  $\sigma_m^2$  quantify the spread of the runs from each model in the historical period, that is, internal variability. The coefficients  $\varphi_m^2$  allow the internal variability of each model to change in the future period. We assume that the historical and future periods are sufficiently separated in time that departures due to internal variability can be considered independent between periods.

To satisfy the assumption of exchangeability between the model outputs, we must also specify the same probability model for the expected climate of each model  $X_{Hm}$  and  $X_{Fm}$ , and the internal variability of each model  $\sigma_m^2$  and  $\varphi_m^2$ . We model the expected climates as

$$X_{Hm} \sim N(\mu_H, \sigma_H^2) \\ X_{Fm}|X_{Hm} \sim N(\mu_F + \beta(X_{Hm} - \mu_H), \sigma_{F|H}^2) \quad (2)$$

and the internal variabilities as

$$\sigma_m^2 \sim \text{Inv-gamma}\left(\frac{\nu_H}{2}, \frac{\nu_H \psi^2}{2}\right) \\ \varphi_m^2 \sim \text{Inv-gamma}\left(\frac{\nu_F}{2}, \frac{\nu_F \theta^2}{2}\right). \quad (3)$$

The model specific parameters  $X_{Hm}$ ,  $X_{Fm}$ ,  $\sigma_m^2$ , and  $\varphi_m^2$  are assumed to be independent between models  $m$ , conditional on the other parameters. The common means  $\mu_H$  and  $\mu_F$  in Equation (2) are interpreted as the representative climate of the ensemble in the historical and future periods, respectively, that is, representative in the sense that they summarize the climates simulated by the models. The variances  $\sigma_H^2$  and  $\sigma_{F|H}^2$  quantify the spread of the models around the representative climate, that is, model uncertainty. The parameterization of the internal variabilities in Equation (3) implies that  $\psi^2 = 1/E[\sigma_m^{-2}]$ ,  $\theta^2 = 1/E[\varphi_m^{-2}]$ , and  $\theta^2 \psi^2 = 1/E[(\varphi_m \sigma_m)^{-2}]$ . Therefore,  $\psi^2$  and  $\theta^2$  can be interpreted as the representative internal variability of the ensemble. The degrees-of-freedom  $\nu_H$  and  $\nu_F$  control the precision of  $\sigma_m^2$  and  $\varphi_m^2$  and quantify model uncertainty about the internal variability.

The parameter  $\beta$  is intended to capture any linear association between the historical climates and future climate responses of the models, that is, any emergent relationship, and is referred to as the *emergent constraint*. The emergent constraint applies to the expected climates of the models, not the individual runs, because emergent relationships are the result of model/process differences, not internal variability. A value of  $\beta = 1$  implies conditional independence of the expected response  $X_{Fm} - X_{Hm}$  of model  $m$  from its expected historical state  $X_{Hm}$ , that is,  $E[X_{Fm} - X_{Hm}|X_{Hm}] = \mu_F - \mu_H$  for all  $m$ . Any value of  $\beta \neq 1$  implies that the expected historical climate  $X_{Hm}$  is informative for the expected climate response  $X_{Fm} - X_{Hm}$ .

The representation in terms of the common unknown means,  $\mu_H$  and  $\mu_F$ , induces a common prior correlation (dependence) between the model means and consequently the model outputs, for example,  $\text{cov}(X_{Hm}, X_{Hm'}) = \text{var}(\mu_H)$  for all  $m \neq m'$ . Thus, we do not require the much stronger assumption that the model outputs are independent.

### 3.2. The Earth System

Let  $Y_{ta}$  represent the single actualization of the Earth system that we observe during time period  $t$ . We model the actualized climate as

$$Y_{Ha}|Y_H \sim N(Y_H, \sigma_a^2) \\ Y_{Fa}|Y_F \sim N(Y_F, (\varphi_a \sigma_a)^2). \quad (4)$$

The means  $Y_H$  and  $Y_F$  represent the expected climate of the Earth system in the historical and future periods, respectively. The variance  $\sigma_a^2$  quantifies the historical natural variability in the Earth system, and the coefficient  $\varphi_a^2$  represents any future change in variability.

Since each model attempts to approximate the Earth system as realistically as possible, we should hope that both the expected climate and internal variability simulated by each model is informative climate of the Earth system. While there are differences between individual climate models, both the mean climates and internal variabilities are usually similar to the Earth system, that is, the observed quantities usually (but not always) lie within the range simulated by the different models. We model the expected climate and natural variability of the Earth system conditional on the representative model as

$$\begin{aligned} Y_H &\sim N(\mu_H, \sigma_{\Delta H}^2) \\ Y_F|Y_H &\sim N(\mu_F + \beta(Y_H - \mu_H), \sigma_{\Delta FH}^2) \end{aligned} \quad (5)$$

and

$$\begin{aligned} \sigma_a^2 &\sim \text{Inv-gamma}\left(\frac{\nu_{Ha}}{2}, \frac{\nu_{Ha}\psi^2}{2}\right) \\ \varphi_a^2 &\sim \text{Inv-gamma}\left(\frac{\nu_{Fa}}{2}, \frac{\nu_{Fa}\theta^2}{2}\right). \end{aligned} \quad (6)$$

In Equation (5), we assume that the emergent constraint  $\beta$  has a well understood physical basis, and therefore applies to the Earth system in the same way as the climate models. The variances  $\sigma_{\Delta H}^2$  and  $\sigma_{\Delta FH}^2$  quantify our uncertainty about the effects of common differences between the models and the Earth system, that is, model inadequacy. Equation (6) implies that  $E[1/\sigma_a^2] = 1/\psi^2$  and  $E[1/\varphi_a^2] = 1/\theta^2$ . The degrees-of-freedom  $\nu_{Ha}$  and  $\nu_{Fa}$  quantify model inadequacy in simulating natural variability in the Earth system. In the language of Rougier, Goldstein, and House (2013), the Earth system is assumed to be *coexchangeable* with the models. Conditioning on the representative model induces a correlation (dependence) between the expected climate and the model means, that is,  $\text{cov}(Y_H, X_{Hm}) = \text{var}(\mu_H)$  for all  $m$ .

### 3.3. The Observed Climate

Let  $Z_H$  be the observed climate during the historical period. We model the observed climate as

$$Z_H \sim N(Y_{Ha}, \sigma_Z^2). \quad (7)$$

The variance  $\sigma_Z^2$  quantifies our observation uncertainty.

### 3.4. Making Inferences About Future Climate

The multi-model ensemble is described by nine parameters  $\mu_H$ ,  $\mu_F$ ,  $\beta$ ,  $\sigma_H^2$ ,  $\sigma_{F|H}^2$ ,  $\psi^2$ ,  $\theta^2$ ,  $\nu_H$ , and  $\nu_F$ . Given outputs from a moderate number of climate models  $M$ , it should be possible to obtain reasonable inferences for the mean parameters  $\mu_H$ ,  $\mu_F$ , and  $\beta$ , and the model uncertainty  $\sigma_H^2$  and  $\sigma_{F|H}^2$ . The internal variability  $\psi^2$  and  $\theta^2$  can be distinguished from model uncertainty provided we have multiple initial condition runs from

several models. Some models may have only a single initial condition run in one or both time periods. In that case, our hierarchical framework allows the model specific internal variability  $\sigma_m^2$  and  $\varphi_m^2$  to be estimated by borrowing strength from models with multiple runs, under the assumption that models should have similar internal variability (Equation (3)). The most difficult parameters to infer are likely to be the degrees-of-freedom  $\nu_H$  and  $\nu_F$ , since these are essentially variances of variances.

The Earth system is represented by a further four parameters  $\sigma_{\Delta H}^2$ ,  $\sigma_{\Delta FH}^2$ ,  $\nu_{Ha}$ , and  $\nu_{Fa}$ . The future parameters  $\sigma_{\Delta FH}^2$  and  $\nu_{Fa}$  cannot be estimated from data, since we have no future observations of the Earth system. Therefore, additional modeling assumptions are required. If an estimate of the historical natural variability  $\sigma_a^2$  is available, then this can be substituted directly, otherwise it can be inferred from the representative model using Equation (6).

In principle, the historical model inadequacy quantified by  $\sigma_{\Delta H}^2$  and  $\nu_{Ha}$  could be estimated from a time series of observations and corresponding simulations. This would require careful modeling to account for time-varying trends and to separate model inadequacy from internal variability and natural variability. In addition, an *extremely* long time series would be required, since the discrepancy between the Earth system and the ensemble is expected to change only slowly over time. Instead, we adopt the approach proposed by Rougier, Goldstein, and House (2013) and parameterize the model inadequacy as proportional to the ensemble spread

$$\begin{aligned} \sigma_{\Delta H}^2 &= \kappa^2 \sigma_H^2 & \sigma_{\Delta FH}^2 &= \kappa^2 \sigma_{F|H}^2, \\ \nu_{Ha} &= \nu_H / \kappa^2 & \nu_{Fa} &= \nu_F / \kappa^2, \end{aligned} \quad (8)$$

where  $\kappa \geq 1$ . The coefficient  $\kappa$  acts to inflate the ensemble spread to account for uncertainty due to processes not well captured by any model, and errors common to all models.  $\kappa$  can be interpreted as quantifying how much less informative the representative model is for the Earth system than for a new climate model comparable to those already in the ensemble. Setting  $\kappa = 1$  implies that the Earth system is exchangeable with the climate models, that is, just another computer model. The value of  $\kappa$  must be fixed a priori, and Rougier, Goldstein, and House (2013) suggested a value of  $\kappa = 1.2$  for surface temperature. Larger values of  $\kappa$  might be appropriate for less well simulated processes, for example, when the models are less informative for the real world and the observations lie outside of the spread in the models.

## 4. Discussion

The hierarchical framework proposed in Section 3 is an extension of the coexchangeable framework of Rougier, Goldstein, and House (2013) to account for internal variability in the models and natural variability in the Earth system. Our framework also makes the role of emergent constraints in constraining model inadequacy explicit.

It is well known that errors in the independent variable in a regression will cause the slope estimate to be biased toward zero, a phenomenon known as *regression dilution* or *regression attenuation* (Frost and Thompson 2000). Therefore, frameworks that ignore internal variability, such as those proposed by Bracegirdle and Stephenson (2012) and Bowman et al. (2018), risk

biased estimates of emergent constraints and hence biased projections. Like the earlier frameworks proposed by Tebaldi et al. (2005), Smith et al. (2009), and others, these frameworks also ignore model inadequacy. This makes them unrealistic unless we believe there are no missing processes in the climate models and that discretizing space and time does not affect our projections. A more detailed comparison between the framework proposed here and these and other existing methods for analyzing multi-model ensemble experiments and emergent constraints is provided in the supplementary materials.

A variety of model weighting schemes have been proposed in the literature, a number of examples are given in Section 1. In principle, model weighting will respect emergent relationships. Consider the example in Figure 1. 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, the weights 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.

The framework proposed here was developed to model temperature data for which the normal distribution is a natural choice. However, since the model outputs  $X_{tmr}$  are assumed to be time-averages, for example, 30-year means, the central limit theorem guarantees that the distribution of the  $X_{tmr}$  should converge to a normal distribution, regardless of the underlying distribution. Therefore, the proposed framework should be suitable for a wide range of other climate variables. If necessary, different distributional choices can be substituted provided the hierarchical structure is respected to maintain the assumption of exchangeability between the model outputs.

In our application, the historical and future variables are the same, that is, temperature. However, there are many examples of emergent relationships in the literature between different variables in the historical and future periods, for example, Cox, Huntingford, and Williamson (2018) related historical temperature variability to equilibrium climate sensitivity. The framework proposed in Section 3 is easily generalized to the case of different historical and future variables by making the future internal variability independent of the historical internal variability in Equation (1), and likewise the natural variability in Equation (4), that is,  $\text{var}(X_{Fmr}) = \varphi_m^2$  rather than  $\text{var}(X_{Fmr}) = \varphi_m^2 \sigma_m^2$ . No other changes are necessary since all other quantities are specified independently for historical and future variables.

The formulation of the emergent relationship in Equations (2) and (5) reflects the linear relationships that have so far been documented in the literature. Bracegirdle and Stephenson (2012) also considered quadratic relationships and Hall et al. (2019) proposed the existence of more general functional relationships. The methodology proposed here generalizes immediately to polynomial relationships and could easily be generalized to other parametric forms.

In Equations (2) and (3), we assume that the climate models are exchangeable, that is, they can be considered independent conditional on the representative model. If we treat models that share common components as independent then we risk unfairly weighting particular groups of models. Methods

for assessing model dependence based on comparing spatial-temporal outputs have been shown to successfully capture similarities between groups of related models (Masson and Knutti 2011; Knutti, Masson, and Gettelman 2013). However, current methods lack a formal statistical framework for combining projections from different models, and can produce unexpected results where models that are known to have little in common are considered close (Sanderson, Knutti, and Caldwell 2015b). Rougier, Goldstein, and House (2013) addressed the problem of model dependence by selecting a subset of models that they judge a priori to be exchangeable. We adopt a similar approach in Section 5 based on readily available data about climate model structure and components shared between models. By analyzing only a subset of the available data we risk losing valuable information. However, the information loss is likely to be acceptable given the known similarities between many climate models (Annan and Hargreaves 2011; Pennell and Reichler 2011)

The framework proposed here makes no assumptions about spatial dependence. Climate model output is often analyzed grid box by grid box, and this is the approach we take in Section 5. In practice, nonphysical discontinuities between neighboring grid boxes are rarely a problem due to the inherent smoothness of computer model output in comparison to observations. Accounting for spatial dependence could potentially lead to more efficient estimates by borrowing strength across neighboring grid boxes. However, any increase in efficiency would come at the cost of additional complexity both in terms of the number of parameters and the computational requirements of fitting to all grid boxes simultaneously.

In the framework proposed here, we adopt the approach introduced by Chandler (2013) and Rougier, Goldstein, and House (2013) and represent multi-model inadequacy as an unknown discrepancy between the climate system and a representative model. This generalizes the well-established single model approach in the uncertainty quantification literature (Craig et al. 2001; Kennedy and O'Hagan 2001) by splitting the discrepancy into two parts: one common to all models, and one unique to each model. The limitations of climate models in approximating the Earth system may manifest themselves in a variety of ways. In the absence of stronger beliefs about how these limitations will manifest, an unknown discrepancy is the simplest and most intuitive way of representing the possibility. Other approaches to representing model inadequacy in an ensemble of computer models may be possible, but we are not aware of any published alternatives.

In Section 1, we made the distinction between a purely statistical “emergent relationship,” and an “emergent constraint” for which a plausible physical mechanism has been identified. Hall et al. (2019) made a similar distinction between what they call “proposed” and “confirmed” emergent constraints, and outline how a constraint might transition from “proposed” to “confirmed.” In formulating our framework, we assume that the emergent constraint applies in the Earth system the same way it does in the models. This assumption is implicit in *all* projections based on emergent constraints, although never stated. By formulating a principled statistical framework, we make this assumption clear and transparent. Thus, by making a projection based on an emergent relationship, we are making a strong statement of confidence in that relationship. The framework

proposed here addresses this by separating model inadequacy from model uncertainty, that is, by allowing for additional uncertainty about the response of the Earth system. However, the appropriate amount of additional uncertainty remains a subjective choice.

### 5. Improved Estimates of Arctic Climate Change

The CMIP5 ensemble includes output from more than 40 models submitted by over 20 centers around the world. To satisfy the assumption of exchangeability in Section 3, we consider a subset of the models that we judge to be approximately exchangeable. The thinned ensemble consists of 13 models and includes 50 runs of the historical period and 39 runs of the future period under the RCP4.5 emissions scenario. The models included in the thinned ensemble were chosen to have similar horizontal and vertical resolutions, but to minimize common component models. In particular, only one model was retained from any one modeling center, usually the most recent and feature complete version submitted by each center. Full details of the thinning process, the included models, and the number of runs from each model are given in the supplementary materials. Our approach to ensemble thinning differs from that of Rougier, Goldstein, and House (2013) who chose models judged to be most similar to a familiar model, effectively minimizing the differences between the models. In contrast, by choosing models with the fewest common components, we are effectively maximizing the differences between the models. In doing so, we aim to capture the broadest range of uncertainty due to model differences. For consistency, we adopt the assessment made by Rougier, Goldstein, and House (2013) and set  $\kappa = 1.2$ . Observation uncertainty  $\sigma_Z^2$  was estimated by combining several different observational datasets. Posterior analysis was performed for each grid box separately. Identical conjugate prior distributions were specified at all grid boxes. Posterior inference proceeds by Gibbs' sampling with Metropolis-Hastings steps for  $\nu_H$  and  $\nu_F$ . Full details of the prior specifications, posterior computation and how we estimate observation uncertainty are given in the supplementary materials.

### 5.1. Model Checking

We checked the assumption of exchangeability between models using a leave-one-out cross-validation approach similar to Smith et al. (2009) and Rougier, Goldstein, and House (2013). Each model in turn is left out of the analysis, and the expected response  $X_{Fm}^* - X_{Hm}^*$  of a new model is predicted. The predictions are compared to the model output using a probability integral transform, that is, by computing the probability that the response under the leave-one-out predicted distribution is less than the mean response of the excluded model. If the models are exchangeable, then the distribution over the models of the transformed projections should be uniform. Kolmogorov–Smirnov tests were used to assess uniformity at each grid box. A small amount of nonuniformity is expected due to shrinkage of the representative climate toward the observations. First, we withheld all data from each model in turn. There was no evidence against the null hypothesis that the models are exchangeable at the 10% level. Second, we withheld only the future simulations to test conditional exchangeability given any emergent relationships. Only two grid boxes were significantly nonuniform at the 10% level. The cross-validation procedure suggests that the chosen models can be considered exchangeable.

### 5.2. Results

The posterior mean estimates of the expected historical climate  $Y_H$ , future climate  $Y_F$ , and climate response  $Y_F - Y_H$  are shown in Figure 3. The 0 °C contour that approximates the sea ice edge has receded noticeably in the projected future climate  $Y_F$  in Figure 3(b) compared to the historical climate  $Y_H$  in Figure 3(a). The projected warming tends to increase with latitude in Figure 3(c).

Figure 4 shows the effects of emergent relationships in near-surface temperature in the Arctic. The posterior mean estimate of the historical discrepancy between the expected climate  $Y_H$  and the representative climate  $\mu_H$  is 2 °C–3 °C across most of the Arctic (Figure 4(a)). The historical discrepancy is largest in the Greenland and Barents seas. This may be due to differences in ocean heat transport simulated by the models (Mahlstein

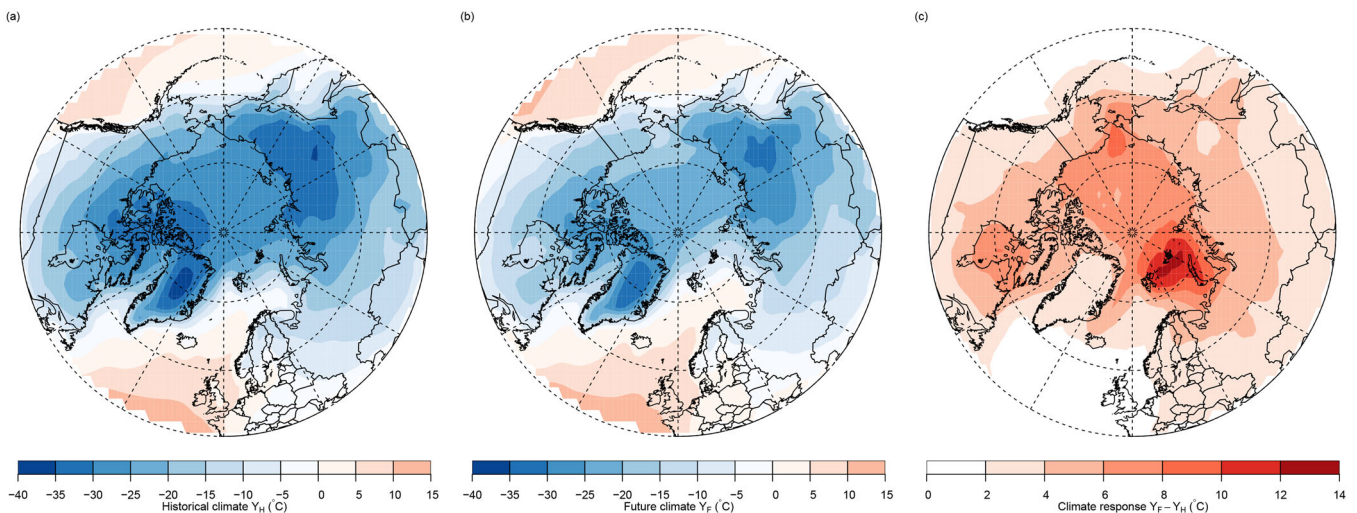
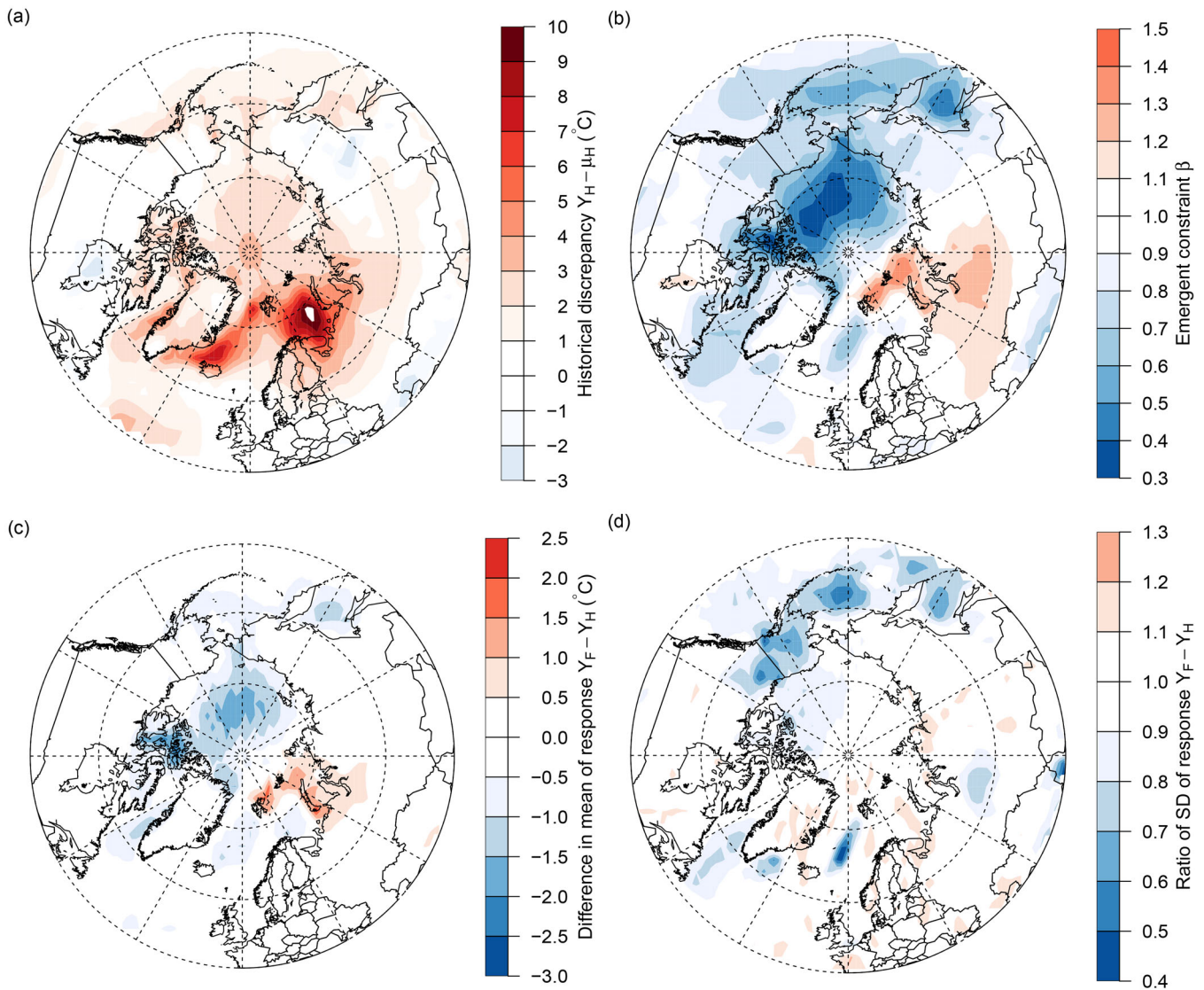


Figure 3. Expected climate. The posterior mean of (a) the historical climate  $Y_H$ ; (b) the future climate  $Y_F$ ; and (c) the climate response  $Y_F - Y_H$ .





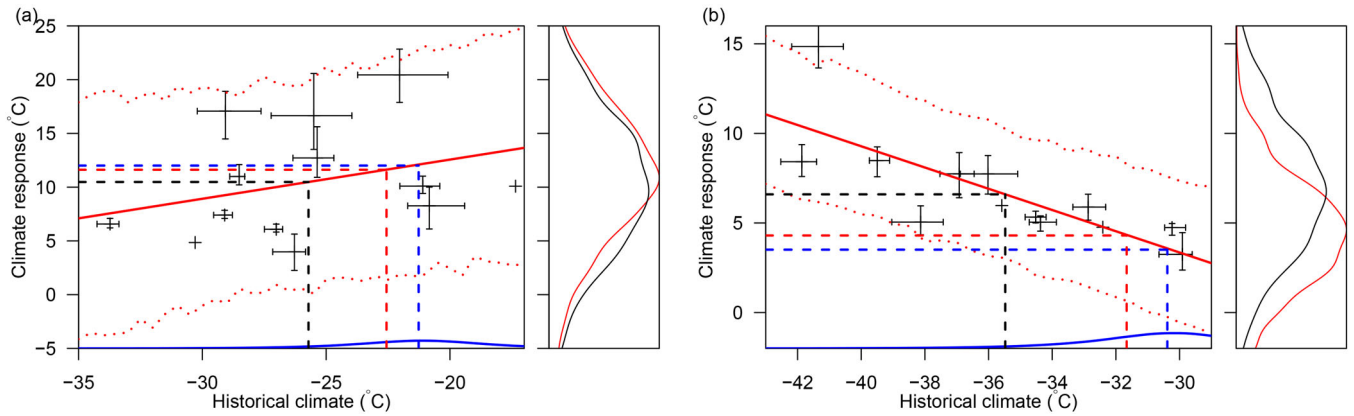
**Figure 4.** Effect of emergent constraints. The posterior mean of (a) the historical discrepancy  $Y_H - \mu_H$ , (b) the emergent constraint  $\beta$ , and (c) the difference in the projected climate response  $Y_H$  due to the emergent constraint  $(\beta - 1)(Y_H - \mu_H)$ . (d) Ratio of posterior standard deviation of the response  $Y_F - Y_H$  with and without an emergent constraint.

and Knutti 2011). From Equation (5), the expected climate response is  $E[Y_F - Y_H|Y_H] = \mu_F - \mu_H + (\beta - 1)(Y_H - \mu_H)$ . The difference in the projected response due to the emergent relationship is given by  $(\beta - 1)(Y_H - \mu_H)$  and is plotted in Figure 4(c). The expected warming is reduced by up to  $3^\circ\text{C}$  in the far north of Canada, and by around  $1^\circ\text{C}$  along most of the ice edge. Figure 4(d) compares the posterior uncertainty about the climate response  $Y_F - Y_H$  with and without an emergent constraint. Around the ice edge, the emergent constraint reduces the posterior standard deviation of the climate response  $Y_F - Y_H$  by 20%–30%.

Our posterior mean estimate of the emergent relationship in the Beaufort sea, north of Alaska in Figure 4(b), is much greater than that of Bracegirdle and Stephenson (2013). Internal variability is small compared to model uncertainty in the Arctic (not shown), so the difference is not due to regression dilution in the ensemble regression estimates. Bracegirdle and Stephenson (2013) analyzed an ensemble of 22 CMIP5 models, some of which were excluded from the ensemble analyzed here. Further

investigation revealed that two of the models excluded from our analysis are strongly warm biased in this region, and two are strongly cold biased, but all four simulate similar climate responses. This acts to neutralize the emergent relationship evident in the remaining models (not shown) in the analysis of Bracegirdle and Stephenson (2013).

The greatest warming occurs near the islands of Svalbard and Franz Josef Land in the north of the Barents sea. Figure 5(a) investigates the strong warming near Svalbard in detail. The representative climate response  $\mu_F - \mu_H$  in Figure 5(a) is already high at  $10.5^\circ\text{C}$  (90% equal-tailed credible interval  $7.7^\circ\text{C}$ – $13.3^\circ\text{C}$ ). The representative response may be influenced by 3 models with unusually large responses. There is a positive emergent relationship  $\beta = 1.4$  (0.8, 2.0) at this grid box, and a historical discrepancy of  $Y_H - \mu_H = 3.0^\circ\text{C}$  ( $-2.7^\circ\text{C}$ ,  $8.2^\circ\text{C}$ ). The emergent relationship predicts an additional  $1.1^\circ\text{C}$  ( $-1.6^\circ\text{C}$ ,  $4.8^\circ\text{C}$ ) of warming. This is relatively insignificant compared to the uncertainty about the response, even when conditioned on the historical climate. The emergent



**Figure 5.** Gridbox details. Data and projections from grid boxes (a) north of Svalbard ( $81^{\circ}\text{N}, 39^{\circ}\text{E}$ ), and (b) east of Devon Island ( $76^{\circ}\text{N}, 94^{\circ}\text{W}$ ). The solid red line indicates the estimated emergent relationship and the dotted red lines indicate a 90% credible interval. The black dashed line indicates the representative climate  $\mu_H$  and climate response  $\mu_F - \mu_H$ . The red dashed line indicates the expected climate  $Y_H$  and climate response  $Y_F - Y_H$ . The blue density represents the distribution of the observations. The blue dashed line indicates the observed climate  $Z_H$  and the climate response based directly on the observations. Auxiliary plots in the right hand margins show the posterior distribution of the climate response  $Y_F - Y_H$  with (red) and without (black) an emergent constraint.

relationship here does little to constrain our uncertainty about the climate response.

Bracegirdle and Stephenson (2013) also estimated a positive emergent relationship over Svalbard, Franz Josef Land, and parts of Siberia, similar to that in Figure 4(b). The posterior probability that  $\beta > 1$  exceeds 0.90 over Western Siberia. Emergent constraints in air temperatures over polar land regions are particularly relevant for constraining estimates of changes in permafrost, which by melting in future could lead to accelerated emissions in greenhouse gases such as methane (Burke, Jones, and Koven 2013). There are significant differences in model temperatures over polar land regions related to model representation of processes such as snow physics and soil hydrology (Koven, Riley, and Stern 2013; Slater and Lawrence 2013). It remains an interesting open question as to why models are showing a positive emergent relationship in the vicinity Western Siberia.

In contrast, a negative emergent relationship is visible in the North West Passage near Devon Island in northern Canada in Figure 5(b). The representative climate response  $\mu_F - \mu_H$  in Figure 5(b) is more moderate at  $6.6^{\circ}\text{C}$  ( $5.1^{\circ}\text{C}$ ,  $8.0^{\circ}\text{C}$ ). There is a negative emergent relationship  $\beta = 0.4$  ( $0.2, 0.7$ ) and a historical discrepancy of  $Y_H - \mu_H = 3.7^{\circ}\text{C}$  ( $-1.1^{\circ}\text{C}$ ,  $8.4^{\circ}\text{C}$ ). The emergent relationship combines with the historical discrepancy to project  $2.2^{\circ}\text{C}$  ( $-5.3^{\circ}\text{C}$ ,  $0.6^{\circ}\text{C}$ ) less warming than the representative model. At this grid box, our uncertainty is usefully constrained by the emergent relationship. The modification to both the mean and standard deviation of the posterior projected response is shown in the right hand margin of Figure 5(b). The posterior standard deviation of the projected response  $Y_F - Y_H$  is reduced by 18%, falling from  $3.9^{\circ}\text{C}$  to  $3.2^{\circ}\text{C}$ .

The examples of Svalbard and Devon Island in Figure 5 both demonstrate the important role of observation and sampling uncertainty when combining models and observations. Due to the sparsity of observations in these remote regions, the observation uncertainty is quite large relative to the model uncertainty. In both cases, there is noticeable shrinkage of the posterior mean estimate of the historical climate  $Y_H$  away from the observations  $Z_H$  and toward the representative climate  $\mu_H$ . As a result, the projected response  $Y_F - Y_H$  lies closer to the

representative response  $\mu_F - \mu_H$  than it would if observation uncertainty were ignored.

## 6. Conclusion

Emergent relationships have become an important topic in climate science for their potential to constrain our uncertainty about future climate change. In this study, we have argued that such relationships can be used to constrain discrepancies due to model inadequacy, if a physical mechanism for the relationship can be identified. The negative emergent constraint on near surface temperature in the Arctic is well understood, and our analysis broadly confirms the findings of previous studies. The projected warming in the Arctic is reduced by up to  $3^{\circ}\text{C}$  by the emergent constraint. Internal variability in the Arctic is large compared to lower latitudes, but is dwarfed by model uncertainty due to the difficulty of representing the many complex processes involved in simulating sea ice, snow cover and the polar vortex. Therefore, regression dilution is unlikely to have significantly biased previous studies of Arctic climate change. However, the sparsity of observations in the Arctic means there is significant observation uncertainty, and this is the first time that observation uncertainty has been accounted for when exploiting emergent constraints. Shrinkage of the expected climate toward the representative climate results in differences of up to  $1^{\circ}\text{C}$  in the projected response compared to estimates based on the observations directly.

The main contribution of this study is to link the concepts of model inadequacy in an ensemble of models and emergent relationships. The proposed Bayesian hierarchical framework also allows the inclusion of multiple runs from each simulator for the first time in a practical application. This allows us to separate uncertainty due to differences between models from internal variability within models. It is differences in the representation of key processes that lead to emergent relationships. Initial conditions should be forgotten over sufficiently long time scales, and therefore should not lead to emergent behavior. We have shown that if internal variability

is not accounted for, then projections based on emergent constraints may be biased. Future multi-model studies exploiting emergent constraints should include multiple runs from each simulator to separate model uncertainty from internal variability and avoid potentially biased projections. Another unique aspect of the framework proposed here is the separation of natural variability and observation uncertainty in the climate system.

The framework proposed in this study allows robust estimation and projection using emergent constraints, but there are still open problems to be addressed both in general multi-model experiments and emergent relationships. The methodology proposed here allows projection of time mean climate accounting for uncertainty due to natural variability. If time-series realizations of natural variability are required within the future study period, for example, for adaptation studies, then our methodology could be extended using the time-series approach proposed by Tebaldi and Sansó (2009), or by transforming observations as proposed by Poppick et al. (2016). Where emergent constraints have been studied at a local level, rather than an aggregate or process level, they have been analyzed one grid box at a time. Ignoring spatial dependence between grid boxes may lead to overly smooth estimates, due to differences in feature placement between models. To obtain physically realistic inferences, spatial statistical methods are required that can represent spatial dependence while accounting for differences between models.

The CMIP5 multi-model ensemble included four future emissions scenarios, but we have analyzed only one. Like climate models, emissions scenarios are difficult to interpret together as an ensemble. Innovative methods are required to extract meaningful probabilistic projections that span the likely range of future emissions. Another source of uncertainty not usually addressed in multi-model experiments is uncertainty about the internal parameters of the climate models. The computational cost of running large perturbed-parameter ensembles is prohibitive. However, each model undergoes a tuning process during which the internal parameters are tested and fixed. Statistical emulators for key quantities, trained during this tuning process, might provide a way of integrating parameter uncertainty into multi-model experiments to provide a more holistic assessment of our uncertainty.

To satisfy the assumption of exchangeability we analyze only a subset of the available models. By adopting this approach, we risk losing valuable information contained in runs from other models and ignoring more detailed insights about model dependence that could be gained by comparing model outputs. In principle, additional levels could be added to the hierarchy proposed here to represent models that share components or were built by the same group. However, the complex overlapping relationships make such a highly structured approach problematic. Current methods for quantifying model dependence based on comparing spatial-temporal output patterns ignore all the prior knowledge we have about the relationships between models. One way forward might be to develop frameworks that combine grouping based on comparing spatial-temporal outputs with simple judgments based on prior knowledge of model inter-dependence. Until alternative methods are found, we recommend thinning the ensemble to

obtain an approximately exchangeable set of models and transparently documenting the thinning process. This does require some prior knowledge on the part of the analyst. However, the burden could be alleviated by establishing standard lists of models, for example, centers submitting to model inter-comparison projects could be asked to nominate a primary model for analysis. This opens up the interesting question of multi-model experiment design. However, the greatest statistical challenge in climate projection is meaningful quantification of model inadequacy. The results here and in Rougier, Goldstein, and House (2013) demonstrate how far we can go with simple judgments. Specifying the model inadequacy via the coefficient  $\kappa$  forces the analyst to make a transparent statement about how informative they believe the models are for the Earth system. However, additional co-operation between statisticians and climate scientists is required to make further progress.

## Supplementary Materials

The online supplementary materials include an extended theoretical comparison with existing multi-model frameworks, a full description of the ensemble thinning process and the included models and runs, full details of our approach to estimating observation uncertainty, the derivation of the Gibbs-Metropolis updating equations, details of the posterior sampling and checking procedures, and plots of additional posterior parameter estimates for the representative climate and the observations.

The data and code used in this study are available from <https://doi.org/10.5281/zenodo.4279112>.

## Acknowledgments

The authors thank Stefan Siegert and Daniel Williamson for helpful comments and discussions

## Funding

This work was supported by the Natural Environment Research Council grant NE/I00520X/1.

## References

- Allen, M. R., and Ingram, W. J. (2002), "Constraints on Future Changes in Climate and the Hydrologic Cycle," *Nature*, 419, 224–232. [2]
- Annan, J. D., and Hargreaves, J. C. (2010), "Reliability of the CMIP3 Ensemble," *Geophysical Research Letters*, 37, L02703. [2]
- (2011), "Understanding the CMIP3 Multimodel Ensemble," *Journal of Climate*, 24, 4529–4538. [2,6]
- Bhat, K. S., Haran, M., Terando, A., and Keller, K. (2011), "Climate Projections Using Bayesian Model Averaging and Space–Time Dependence," *Journal of Agricultural, Biological, and Environmental Statistics*, 16, 606–628. [2,6]
- Bishop, C. H., and Abramowitz, G. (2013), "Climate Model Dependence and the Replicate Earth Paradigm," *Climate Dynamics*, 41, 885–900. [2,3]
- 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, 13050–13059. [3,5]
- 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, 2805–2821. [2,3,5,6]
- (2013), "On the Robustness of Emergent Constraints Used in Multimodel Climate Change Projections of Arctic Warming," *Journal of Climate*, 26, 669–678. [3,8,9]

- Brient, F. (2020), "Reducing Uncertainties in Climate Projections With Emergent Constraints: Concepts, Examples and Prospects," *Advances in Atmospheric Sciences*, 37, 1–15. [2]
- 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, 4897–4909. [9]
- 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, 849–868. [2]
- 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. [1,2,6]
- 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, 1957–1970. [1]
- 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, 403–409. [2]
- Cox, P. M., Huntingford, C., and Williamson, M. S. (2018), "Emergent Constraint on Equilibrium Climate Sensitivity From Global Temperature Variability," *Nature*, 553, 319–322. [2,3,6]
- 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, 717–729. [1,2,6]
- 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 *Climate Change 2001: The Scientific Bases. Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change*, eds. J. Houghton, Y. Ding, D. J. Griggs, M. Noguer, P. J. van der Linden, K. Dai, X. Maskell, and C. A. Johnson, Cambridge: Cambridge University Press, p. 881. [2]
- 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. [3]
- 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*, 163, 173–189. [5]
- 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, 249–266. [2]
- Greene, A. M., Goddard, L., and Lall, U. (2006), "Probabilistic Multimodel Regional Temperature Change Projections," *Journal of Climate*, 19, 4326–4343. [2]
- Hall, A., Cox, P., Huntingford, C., and Klein, S. (2019), "Progressing Emergent Constraints on Future Climate Change," *Nature Climate Change*, 9, 269–278. [2,6]
- 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, 1–4. [3]
- 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. [3]
- (2011), "The Potential to Narrow Uncertainty in Projections of Regional Precipitation Change," *Climate Dynamics*, 37, 407–418. [3]
- Holland, M. M., and Bitz, C. M. (2003), "Polar Amplification of Climate Change in Coupled Models," *Climate Dynamics*, 21, 221–232. [3]
- 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, 934–947. [2]
- Kennedy, M. C., and O'Hagan, A. (2001), "Bayesian Calibration of Computer Models," *Journal of the Royal Statistical Society, Series B*, 63, 425–464. [2,6]
- 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, 2739–2758. [1,2]
- Knutti, R., Masson, D., and Gettelman, A. (2013), "Climate Model Genealogy: Generation CMIP5 and How We Got There," *Geophysical Research Letters*, 40, 1194–1199. [3,6]
- 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, 1909–1918. [2,6]
- 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, 1877–1900. [9]
- Lambert, S. J., and Boer, G. J. (2001), "CMIP1 Evaluation and Intercomparison of Coupled Climate Models," *Climate Dynamics*, 17, 83–106. [2]
- Mahlstein, I., and Knutti, R. (2011), "Ocean Heat Transport as a Cause for Model Uncertainty in Projected Arctic Warming," *Journal of Climate*, 24, 1451–1460. [8]
- Masson, D., and Knutti, R. (2011), "Climate Model Genealogy," *Geophysical Research Letters*, 38, L08703. [6]
- McKinnon, K. A., and Deser, C. (2018), "Internal Variability and Regional Climate Trends in an Observational Large Ensemble," *Journal of Climate*, 31, 6783–6802. [3]
- 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, L08708. [2]
- 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, 747–756. [1,3]
- Northrop, P. J., and Chandler, R. E. (2014), "Quantifying Sources of Uncertainty in Projections of Future Climate," *Journal of Climate*, 27, 8793–8808. [3]
- Oreskes, N., Shrader-Frechette, K., and Belitz, K. (1994), "Verification, Validation, and Confirmation of Numerical Models in the Earth Sciences," *Science*, 263, 641–646. [2]
- Parker, W. S. (2006), "Understanding Pluralism in Climate Modeling," *Foundations of Science*, 11, 349–368. [1]
- Pennell, C., and Reichler, T. (2011), "On the Effective Number of Climate Models," *Journal of Climate*, 24, 2358–2367. [6]
- Poppick, A., McNerney, 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, 477–505. [10]
- Qu, X., and Hall, A. D. (2014), "On the Persistent Spread in Snow-Albedo Feedback," *Climate Dynamics*, 42, 69–81. [3]
- 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, 3212–3226. [2]
- Rougier, J. C., Goldstein, M., and House, L. (2013), "Second-Order Exchangeability Analysis for Multimodel Ensembles," *Journal of the American Statistical Association*, 108, 852–863. [2,5,6,7,10]
- Sanderson, B. M., Knutti, R., and Caldwell, P. M. (2015a), "A Representative Democracy to Reduce Interdependency in a Multimodel Ensemble," *Journal of Climate*, 28, 5171–5194. [2]
- (2015b), "Addressing Interdependency in a Multimodel Ensemble by Interpolation of Model Properties," *Journal of Climate*, 28, 5150–5170. [2,6]
- 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. [2]
- Slater, A. G., and Lawrence, D. M. (2013), "Diagnosing Present and Future Permafrost From Climate Models," *Journal of Climate*, 26, 5608–5623. [9]
- 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, 97–116. [2,3,6,7]
- 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. [1,2]
- Stephenson, D. B., Collins, M., Rougier, J. C., and Chandler, R. E. (2012), “Statistical Problems in the Probabilistic Prediction of Climate Change,” *Environmetrics*, 23, 364–372. [1,2,3]
- 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, 485–498. [3]
- 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, 2053–2075. [1]
- 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*, 172, 83–106. [2,10]
- 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, 1524–1540. [2,3,6]
- 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, 6443–6456. [3]
- 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, 1–13. [2,3]
- Weigel, A. P., Knutti, R., Liniger, M. A., and Appenzeller, C. (2010), “Risks of Model Weighting in Multimodel Climate Projections,” *Journal of Climate*, 23, 4175–4191. [2]
- 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, 4634–4643. [3]