Simple Uncertainty Frameworks for Selecting Weighting Schemes and Interpreting Multimodel Ensemble Climate Change Experiments

PHILIP G. SANSOM, DAVID B. STEPHENSON, AND CHRISTOPHER A. T. FERRO

University of Exeter, Exeter, United Kingdom

GIUSEPPE ZAPPA AND LEN SHAFFREY

National Centre for Atmospheric Sciences, University of Reading, Reading, United Kingdom

(Manuscript received 22 July 2012, in final form 2 November 2012)

ABSTRACT

Future climate change projections are often derived from ensembles of simulations from multiple global circulation models using heuristic weighting schemes. This study provides a more rigorous justification for this by introducing a nested family of three simple analysis of variance frameworks. Statistical frameworks are essential in order to quantify the uncertainty associated with the estimate of the mean climate change response.

The most general framework yields the ‘‘one model, one vote’’ weighting scheme often used in climate projection. However, a simpler additive framework is found to be preferable when the climate change response is not strongly model dependent. In such situations, the weighted multimodel mean may be interpreted as an estimate of the actual climate response, even in the presence of shared model biases.

Statistical significance tests are derived to choose the most appropriate framework for specific multimodel ensemble data. The framework assumptions are explicit and can be checked using simple tests and graphical techniques. The frameworks can be used to test for evidence of nonzero climate response and to construct confidence intervals for the size of the response.

The methodology is illustrated by application to North Atlantic storm track data from the Coupled Model Intercomparison Project phase 5 (CMIP5) multimodel ensemble. Despite large variations in the historical storm tracks, the cyclone frequency climate change response is not found to be model dependent over most of the region. This gives high confidence in the response estimates. Statistically significant decreases in cyclone frequency are found on the flanks of the North Atlantic storm track and in the Mediterranean basin.

1. Introduction

Future climate projections are usually inferred from simulations from general circulation models. The previous phase of the World Climate Research Programme (WCRP) Coupled Model Intercomparison Project (phase3; CMIP3) included 24 models from 17 groups in 12 countries (Meehl et al. 2007b). The latest CMIP (phase 5; CMIP5) multimodel ensemble (MME) (Taylor et al. 2012) is not yet fully populated but promises to include an even greater number of more recent models (see Table 1 for a full list of models included in this study). These MMEs represent a rich source of data for climate scientists. However, in a recent review, Knutti et al. (2010b) concluded that ‘‘quantitative methods to extract the relevant information and to synthesize it are urgently needed.’’

The models, scenarios, and runs that make up an MME explore the three primary sources of uncertainty in climate projections. Structural (model) uncertainty arises from the fact that not all relevant processes are well represented in models. Different scenarios represent uncertainty about changes in radiative forcing due to future emissions. Ideally, several perturbed initial condition runs of each scenario should also be available from each model in order to sample internal variability. These sources of uncertainty can be quantitatively partitioned using simple analysis of variance (ANOVA) frameworks (Yip et al. 2011).

The projections presented in the Intergovernmental Panel on Climate Change (IPCC) Fourth Assessment Report (Solomon et al. 2007) were largely based on

Corresponding author address: Philip Sansom, Harrison Building, University of Exeter, North Park Road, Exeter EX4 4QF, United Kingdom.
E-mail: pgs201@exeter.ac.uk

DOI: 10.1175/JCLI-D-12-00462.1

TABLE 1. List of CMIP5 models and institutes included in the study.

Modeling center (or group)	Model name	Model expansions
Beijing Climate Center (BCC), China Meteorological Administration	BCC-CSM1.1	BCC Climate System Model, version 1.1
Canadian Centre for Climate Modelling and Analysis	CanESM2	Second-generation Canadian Earth System Model
Centre National de Recherches Météorologiques (CNRM)/Centre Européen de Recherches et de Formation Avancée en Calcul Scientifique	CNRM-CM5	CNRM Coupled Global Climate Model, version 5
Commonwealth Scientific and Industrial Research Organisation (CSIRO) in collaboration with Queensland Climate Change Centre of Excellence	CSIRO-Mk3.6.0	CSIRO, Mark version 3.6.0
EC-Earth consortium	EC-EARTH	
National Key Laboratory of Numerical Modeling for Atmospheric Sciences and Geophysical Fluid Dynamics (LASG), Institute of Atmospheric Physics, Chinese Academy of Sciences, and CESS, Tsinghua University	FGOALS-g2	Flexible Global Ocean–Atmosphere–Land System Model, gridpoint version 2
NOAA/Geophysical Fluid Dynamics Laboratory (GFDL)	GFDL-ESM2G, GFDL-ESM2M	GFDL Earth System Model 2G, GFDL Earth System Model 2G
Met Office Hadley Centre	HadGEM2-CC, HadGEM2-ES	Hadley Centre Global Environment Model, version 2 (Carbon Cycle), Hadley Centre Global Environment Model, version 2 (Earth System)
Institute for Numerical Mathematics (INM)	INM-CM4	INM Coupled Model, version 4
L'Institut Pierre-Simon Laplace (IPSL)	IPSL-CM5A-LR, IPSL-CM5A-MR	IPSL Coupled Model version 5A, low resolution; IPSL Coupled Model version 5A, medium resolution
Atmosphere and Ocean Research Institute (University of Tokyo), National Institute for Environmental Studies, and Japan Agency for Marine-Earth Science and Technology	MIROC5	Model for Interdisciplinary Research on Climate 5
Japan Agency for Marine-Earth Science and Technology, Atmosphere and Ocean Research Institute (University of Tokyo), and National Institute for Environmental Studies	MIROC-ESM, MIROC-ESM-CHEM	Model for Interdisciplinary Research on Climate Earth System Model, Model for Interdisciplinary Research on Climate Earth System Model, atmospheric chemistry coupled version
Max Planck Institute (MPI) for Meteorology	MPI-ESM-LR	MPI Earth System Model, low resolution
Meteorological Research Institute (MRI)	MRI-CGCM3	MRI Coupled General Circulation Model version 3
Norwegian Climate Centre	NorESM1-M	Norwegian Earth System Model 1, medium resolution

arithmetic means of the projections from the models in the CMIP3 MME. When multiple runs are available from a model, these are often averaged together before averaging over all models. Alternatively, one run or a fixed number of runs may be selected from each model for each scenario (e.g., Tebaldi et al. 2011a; Yip et al. 2011). These approaches treat all models as equally credible, that is, "one model, one vote" (Knutti et al. 2010a).

The one model, one vote approach is a point estimate that has a number of shortcomings. The assumptions underlying this heuristic estimate are not explicit and therefore cannot be checked. No assessment of the uncertainty associated with the estimate is given, so confidence intervals on the climate response cannot be constructed. Also, arithmetic means are not resistant estimators and may be strongly influenced by runs that are outliers compared to the rest of the MME.

The shortcomings of the one model, one vote approach may be addressed by specifying our assumptions about the structure of the uncertainty in the MME using a statistical framework. The statistical framework is effectively an emulator for the entire ensemble. If the framework correctly describes the behavior of, for example, the CMIP5 models, then it should be possible to stochastically generate a new ensemble of CMIP5 runs from the statistical framework that would be indistinguishable from the expected result of rerunning the CMIP5 models themselves.

The use of the one model, one vote approach as an estimate of the actual climate response is often justified by the assumption that the model mean climates are centered on the actual climate, that is, ''truth centered'' (Knutti et al. 2010a). However, there is an increasing awareness that GCMs share common biases compared to the actual climate (Knutti et al. 2010b). The existence of common biases is to be expected since climate models are often calibrated against the same data, run at similar resolutions, and share similar numerical codes or entire model components (Stephenson et al. 2012; Collins et al. 2012; Sanderson and Knutti 2012). A number of authors have suggested statistical frameworks that explicitly account for biases between models and the actual climate (Chandler 2013; Rougier et al. 2012, manuscript submitted to *J. Amer. Stat. Assoc.*; Tebaldi et al. 2011b). The common approach in each of these frameworks is to propose a separate statistical model for the relationship among the models in the MME and for the relationship between the MME and the actual climate, linked via the ensemble mean climate.

Some studies (e.g., Giorgi and Mearns 2002; Tebaldi et al. 2005) have weighted models according to how well they simulate past observations and their convergence to the ensemble mean response. In seasonal and interannual climate forecasting, models are often weighted according to performance in simulating observed climate by regressing hindcasts on previous observations (DelSole 2007; Kharin and Zwiers 2002; Peña and van den Dool 2008). The conditions that make these methods attractive over short lead times do not apply over longer lead times (Weigel et al. 2010). If the weights applied do not reflect the true model skill or if the internal variability is large compared to the structural uncertainty, an unweighted estimate may be preferred (Weigel et al. 2010). For these reasons and for the sake of simplicity, this study addresses the prerequisite problem of how to construct suitable estimates in the absence of information about past performance.

This study uses ANOVA frameworks to make explicit one simple set of assumptions that lead naturally to the one model, one vote estimate of the ensemble mean climate response. However, a more precise estimate can be obtained when the structural uncertainty in the climate response is small compared to the internal variability. In that case, it may be possible to neglect the estimation of any shared bias between the models and the actual climate and obtain confidence intervals for the expected actual climate response.

ANOVA frameworks have already been used in climate science for a variety of purposes (Zwiers 1987, 1996; Räisänen 2001). Simple ANOVA frameworks have been used to analyze MMEs of both GCMs (Yip et al. 2011) and regional climate models (RCMs) (Ferro 2004; Hingray et al. 2007). Further studies of MMEs of RCMs have used the ANOVA methodology as the basis for more complex frameworks (Sain et al. 2011; Kang and Cressie 2013).

Section 2 of this paper describes the ANOVA frameworks and their underlying assumptions, methods to verify those assumptions, and a formal statistical approach to choosing which set of assumptions are most appropriate to describe the uncertainty in a particular MME. In section 3, the ANOVA approach is illustrated by application to the future climate response of the North Atlantic storm track in the CMIP5 MME.

2. Statistical frameworks

This section begins with a general discussion of multimodel mean estimates of the climate response in an MME. A family of ANOVA frameworks are then outlined, the most general of which is shown to yield the usual one model, one vote multimodel mean. The rest of the section addresses statistical inference using these frameworks. This includes how to test the underlying assumptions, how to choose the most appropriate framework, and how to construct statistical significance tests and confidence intervals.

a. The multimodel mean response

Let y_{msr} represent a climate statistic (e.g., a 30-yr mean) from run r of scenario s simulated by climate model m. For simplicity, we consider an MME containing only one historical scenario H and one future scenario F. The climate response of model m is usually estimated by the difference between its sample mean climates in the historical and future scenarios

$$\overline{y}_{mF.} - \overline{y}_{mH.}, \qquad (1)$$

where $\overline{y}_{ms.}$ is the sample mean climate simulated by model m in scenario s,

$$\overline{y}_{ms.} = \frac{1}{R_{ms}} \sum_{r=1}^{R_{ms}} y_{msr},$$

and R_{ms} is the number of runs from model m under scenario s. A general multimodel mean estimate of the climate response is given by

$$\frac{1}{W_{.F}} \sum_{m=1}^{M} W_{mF}\overline{y}_{mF.} - \frac{1}{W_{.H}} \sum_{m=1}^{M} W_{mH}\overline{y}_{mH.}, \qquad (2)$$

where

$$W_{.H} = \sum_{m=1}^{M} W_{mH} \quad \text{and} \quad W_{.F} = \sum_{m=1}^{M} W_{mF}$$

and M is the number of models. The W_{mH} and W_{mF} are model specific weights on the historical and future scenarios, respectively. The most commonly used estimate is the equally weighted multimodel mean, that is, the one model, one vote approach, where

$$W_{mH} = W_{mF} = 1 \quad \text{for all models} \quad m = 1, 2, \ldots, M. \tag{3}$$

b. A two-way ANOVA framework with interactions

In the appendix it is shown that the one model, one vote estimate of the climate response from Eq. (3) is equivalent to the maximum-likelihood (ML) estimate $\hat{\beta}_F$ of the ensemble mean climate response from the following two-way ANOVA framework with interactions:

$$\begin{aligned} y_{msr} &= \mu + \alpha_m + \beta_s + \gamma_{ms} + \epsilon_{msr}, \\ \epsilon_{msr} &\overset{\text{iid}}{\sim} N(0, \sigma^2), \end{aligned} \tag{4}$$

with the usual constraints that $\sum_{m=1}^{M}\alpha_m = 0$, $\beta_H = 0$, and $\gamma_{mH} = 0$ for all models and $\sum_{m=1}^{M}\gamma_{mF} = 0$. The effect μ is the expected climate (in the ensemble) in the historical scenario, and β_F is the expected climate response (in the ensemble) to scenario F. The effect α_m is the difference between the mean historical climate of model m and the expected historical climate μ. The interaction terms γ_{mF} represent the difference between the mean climate response simulated by model m and the expected climate response β_F. The constraint $\sum_{m=1}^{M}\alpha_m = 0$ ensures that the mean historical climates of the individual models are centered on the expected historical climate μ. Similarly, the constraint $\sum_{m=1}^{M}\gamma_{mF} = 0$ ensures that the mean climate responses of the individual models are centered on the expected climate response β_F.

The random component ϵ_{msr} represents the internal variability of y_{msr} and is assumed for simplicity to be normally distributed and constant for all models and both scenarios. The central limit theorem implies that any long-term mean will be approximately normally distributed (if the climate response trend is small). The assumption that the internal variability is constant between models is a working assumption and must be checked (see section 2e).

There are a total of $2M$ parameters to be estimated in the ANOVA framework of Eq. (4). One parameter must be estimated for the expected historical climate μ and one for the expected climate response β_F. To avoid ill conditioning, the α_m and γ_{mF} effects are constrained to be centered on μ and β_F, respectively. Therefore, only $M - 1$ of each needs to be estimated. If only two runs of each scenario are available from each model, then there are $N = \sum_m (R_{mH} + R_{mF}) = 4M$ runs in total. If $2M$ degrees of freedom are used up estimating the mean effects, only $2M$ remain to estimate the size of the internal variability σ^2. In a small MME, there is a risk of over fitting, and the precision of the estimates may be low.

If only one run of each scenario is available from each model, then $N = 2M$, and the framework has as many parameters as runs. All the degrees of freedom are then used up estimating the mean effects, and the internal variability represented by the random term ϵ_{msr} cannot be estimated. If the internal variability cannot be estimated, then the framework assumptions cannot be tested, and the significance tests and confidence intervals outlined later in this section cannot be used.

The inclusion of the interaction term γ_{ms} complicates the interpretation of the ensemble expected climate response β_F. If the models all simulate different responses, how can we be confident in how the actual climate will respond? The truth-centered approach assumes that β_F coincides with the actual climate response. However, biases shared by all models mean that this may not be the case (Knutti et al. 2010b).

The α_m and γ_{ms} terms represent the structural (model) uncertainty in the historical climate and climate response, respectively. Their relative contribution to the total uncertainty in the MME is quantified in section 2g. However, only the size of the uncertainty due to internal variability is quantified directly through the ϵ_{msr} terms and the parameter σ^2. Therefore, we do not recommend reporting confidence intervals based on the one model, one vote estimate of the climate response since doing so would neglect the contributions from structural uncertainty and any shared bias. In section 2g it is shown that if the relative contribution of the structural uncertainty in the climate response is sufficiently small compared to the internal variability, we may safely assume $\gamma_{mF} = 0$ for all models, that is, the models simulate the same climate response.

c. A simpler additive ANOVA framework

If the models all simulate the same climate response, then estimating the γ_{mF} effects is unnecessary. Estimating a systematic component where none exists increases variance, which leads to decreased precision in the estimates. More precise estimates may be obtained using a simpler additive framework:

$$\begin{aligned} y_{msr} &= \mu + \alpha_m + \beta_s + \epsilon_{msr}, \\ \epsilon_{msr} &\overset{\text{iid}}{\sim} N(0, \sigma^2), \end{aligned} \tag{5}$$

with the usual constraints that $\sum_{m=1}^{M}\alpha_m = 0$ and $\beta_H = 0$. The effects are interpreted as in the two-way framework of Eq. (4). However, the ML estimates of the effects are not the same.

In the appendix it is shown that the ML estimate $\hat{\beta}_F$ of the expected climate response from the additive framework is a weighted average of the model mean responses with weights

$$W_{mH} = W_{mF} = \frac{R_{mH}R_{mF}}{R_{mH} + R_{mF}}. \tag{6}$$

This additive framework assumes that all models simulate the same climate response with the same internal variability. If that assumption is believable, then we should give increased weight to models that have more runs. This argues against advice to avoid weighting models based on the number of runs they contribute to the MME (Knutti et al. 2010a). Note that the weights depend on the combined number of historical and future runs. To achieve a high weighting, it is necessary to have many runs from both scenarios.

The additive framework is more parsimonious and has only $M + 1$ parameters to be estimated. Without the interaction effects, there are $M - 1$ less parameters to be estimated. An additional $M - 1$ degrees of freedom are then available to estimate the internal variability. Therefore, the precision of the parameter estimates should increase compared to the two-way framework with interactions. However, if the models do not all simulate the same climate response, then a systematic component is missing from the framework. The precision of the estimates will decrease dramatically if the missing effects are large. The additive framework must therefore only be used when the structural uncertainty in the climate response is small compared to the internal variability, as shown in section 2g.

If the models all simulate the same climate response, then no truth-centered assumption is required to justify the mean response of the ensemble as an estimate of the actual climate response. However, the possibility remains of a bias shared by all the models compared to the actual climate. In estimating the actual climate response, any shared bias will cancel if it is constant in both historical and future scenarios. Such an assumption is difficult to verify since we have no observations of the future for comparison. This assumption may still be optimistic (Buser et al. 2009; Christensen et al. 2008); however, it is a more acceptable assumption than that of no shared bias. Therefore, if the conditions outlined in section 2g are satisfied, the only notable uncertainty in the climate response in scenario F is due to internal variability, and the confidence intervals given in the appendix should be reported for the expected climate response β_F.

d. A simple one-way ANOVA framework

The α_m effects allow for the possibility that each model simulates a different historical mean climate. In the unlikely event that all models are believed to simulate the same historical climate, then a one-way ANOVA framework may provide more precise estimates:

$$\begin{aligned} y_{msr} &= \mu + \beta_s + \epsilon_{msr}, \\ \epsilon_{msr} &\overset{\text{iid}}{\sim} N(0,\sigma^2), \end{aligned} \tag{7}$$

with the usual constraint that $\beta_H = 0$. The effects are interpreted as in the more complex frameworks; however, the ML estimates of the effects are not the same.

In the appendix it is shown that the ML estimate $\hat{\beta}_F$ of the expected climate response from this one-way framework is also a weighted average of the model mean responses with weights

$$W_{mH} = R_{mH} \quad \text{and} \quad W_{mF} = R_{mF}. \tag{8}$$

In this case, the weights are equivalent to giving equal weight to every run in the MME, that is, ''one run, one vote.'' Note that in the balanced case where $R_{mH} = R_{mF}$, the weights from the additive framework in Eq. (6) reduce to the one run, one vote estimate.

This simple framework has only two parameters to be estimated. With $M - 1$, additional degrees of freedom available to estimate the internal variability the precision of the estimates should increase again. However, a similar caveat applies as in the additive framework. If the models do not all simulate the same historical climate and climate response, the precision of the estimates may decrease dramatically. The one-way framework must therefore only be used when the structural uncertainty associated with both the historical climate and the climate response is small compared to the internal variability, as shown in section 2g.

The assumptions required in order to justify the one run, one vote estimate as an estimate of the actual climate response are identical to those outlined for the additive framework in section 2c. Therefore, if the conditions outlined in section 2g are satisfied, the confidence intervals given in the appendix should be reported for the expected climate response β_F. However, the estimates will have greater precision compared to those from the additive framework.

e. Is an ANOVA framework appropriate?

The traditional estimation procedure for ANOVA frameworks involves only simple linear combinations of the group means of the various factors included in the framework, that is, the model-scenario means $\overline{y}_{ms.}$. This simplicity comes at the cost of requiring a balanced design, that is, the same number of runs of each model for each scenario. So in an MME, it might be necessary to exclude additional runs from some models, or to exclude models that do not have sufficient runs. This can be avoided by fitting the ANOVA framework using normal linear regression methods (Krzanowski 1998).

There are three main assumptions about the random component in these frameworks:

- the residuals ϵ_{msr} are mutually independent;
- the residuals ϵ_{msr} are normally distributed;
- the residuals ϵ_{msr} have constant variance.

Each of these assumptions must be carefully checked before confidence can be placed in the estimates from the frameworks. If they are satisfied, then the ANOVA framework provides a good statistical description of the MME.

The distributional assumptions may be checked by analysis of the fitted residuals $e_{msr} = y_{msr} - \hat{y}_{msr}$. The fitted values $\hat{y}_{msr}$ from each framework are defined in the appendix. If the data are normally distributed, then a plot of the ordered standardized residuals against the theoretical quantiles of the normal distribution should lie close to a straight line through the origin with unit gradient. If the data have constant variance, then plotting the standardized residuals against the fitted values $\hat{y}_{msr}$ should show random scatter about zero. Any systematic component visible in the scatter may indicate nonconstant variance or a systematic difference between the y_{msr} that is not captured by the framework.

The assumption of independence is less easily checked, so consideration must be given a priori to whether this assumption is justified. Under the truth-centered view, it would be necessary to assume that the model mean climates are distributed independently about the actual climate. However, there is an increasing awareness that this may not be the case (Knutti et al. 2010b). It is less restrictive to assume that the models are independent depending on the ensemble mean climate, that is, independently distributed about the ensemble mean climate (Rougier et al. 2012, manuscript submitted to *J. Amer. Stat. Assoc.*). This splits the model bias into a part that is shared between all models in the MME and a part that is unique to each model, independent of the others. Since we do not consider the actual climate explicitly, we need only consider the independent part. This assumption may still be optimistic (Pennell and Reichler 2011). However, it is a more acceptable assumption than that of complete independence of model biases.

f. Identifying outlying runs

As in any large experiment, there are a variety of ways by which unexpected results may enter into an MME. These include human error (e.g., initialization errors or mislabeling a particular run) as well as less predictable factors (e.g., poorly chosen initial conditions or a parameterization that lacks the flexibility to respond correctly to a particular scenario). The ANOVA frameworks can be used to systematically identify runs that appear to be outliers with respect to the rest of the MME.

The ϵ_{msr} are assumed to be normally distributed. Therefore, fitted residuals e_{msr} should also be normally distributed. Any runs having standardized fitted residuals lying in the far tails of the standard normal distribution are considered outlying. If viewed as a significance test, we might consider labeling any run with a standardized residual in the most extreme 10% of the normal distribution ($|Z| > 1.64$) as outlying. However, the residuals are assumed to be independent, so we would expect 10% of all residuals to lie in this region. A stricter 1% criterion ($|Z| > 2.58$) is therefore more appropriate.

Outliers can be easily identified from the plot of standardized residuals against fitted values $\hat{y}_{msr}$. They may also be visible in the quantile–quantile plot used in the check for normality. As noted above, outlying runs arise for a variety of reasons. They may represent unlikely but still plausible climates and contribute valuable information to the MME. Therefore, outlying runs should not simply be dismissed from the MME unless an explanation can be found for the unusual behavior.

Outlying runs can have a large influence on the parameter estimates. A quick check of the influence of any outliers is to temporarily remove them, refit the framework, and check the parameter estimates. If the estimates of the main effects μ and β_F do not change, then the influence of the outliers is small. In that case, the outlying runs should remain in the ensemble. If removing the outliers strongly affects the estimates of the main effects μ and β_F, then it is essential to determine whether the outlying runs represent plausible climates or problematic simulations.

Outlying runs may also affect the test for normality. A large number of outliers are a strong indication that the framework assumptions are not appropriate. If there are only one or two outliers, then they may simply be results that are unlikely given the total number of runs. This can quickly be checked by temporarily removing the outliers, refitting the framework, and rechecking the normality. If the normality is satisfactory after removing the outliers, then the analysis can proceed with the outlying runs included. If the normality is still not satisfied, an ANOVA framework may not be appropriate.

g. Which framework is most appropriate?

In section 2c it is noted that the additive framework is only appropriate if all models simulate the same climate response. Similarly, in section 2d it is noted that the one-way framework is only appropriate if the models also simulate the same historical climate. These are conditions on model agreement. This is often quantified by the number of models having the same sign of response or discrepancy. That does not take into account the internal variability (Tebaldi et al. 2011a). If the expected climate response β is small compared to the internal variability, then models may appear to disagree when they are actually behaving similarly.

The additive framework is a special case of the two-way framework with interactions where $\gamma_{mF} = 0$ for all models m. In the appendix, a statistical significance test is derived for the presence of model dependence in the climate response, that is, to test the null hypothesis H_0: $\gamma_{mF} = 0$ for all m against the alternative H_a: $\gamma_{mF} \neq 0$ for some m. The test statistic is the ratio of variances:

$$F_\gamma = \frac{N-2M}{M-1} f_\gamma^2, \quad \text{where} \quad f_\gamma^2 = \frac{R_\gamma^2 - R_\alpha^2}{1 - R_\gamma^2}. \tag{9}$$

The statistics R_γ^2 and R_α^2 are the coefficients of determination for the two-way framework with interactions and the additive framework, respectively. The coefficient of determination R^2 is the proportion of total variability explained by a normal linear regression framework. The quantity f_γ^2 therefore represents the ratio of the variance explained by structural uncertainty (model dependence) in the climate response to that explained by internal variability. If the structural uncertainty is small compared to the internal variability, then estimating the γ_{mF} effects does not significantly improve the framework as a description of the MME. Formally, if the p value of the test is small ($p < a$), we conclude that there is significant evidence of model dependence in the climate response at the a% level and that the two-way framework is most appropriate. Otherwise, the additive framework is more appropriate.

Similarly, the one-way framework is a special case of the additive framework where $\alpha_m = 0$ for all models m. In the appendix, a statistical significance test is derived to test for the presence of model dependence in the historical climate, that is, to test the null hypothesis H_0: $\alpha_m = 0$ for all m against the alternative H_a: $\alpha_m \neq 0$ for some m. The test statistic is the ratio of variances:

$$F_\alpha = \frac{N-(M+1)}{M-1} f_\alpha^2, \quad \text{where} \quad f_\alpha^2 = \frac{R_\alpha^2 - R_\beta^2}{1 - R_\alpha^2}. \tag{10}$$

The quantity R_β^2 is the coefficient of determination for the one-way framework. The quantity f_α^2 represents the ratio of the variance explained by structural uncertainty in the historical climate to that explained by internal variability. If the structural uncertainty is small compared to the internal variability, then estimating the α_m effects does not significantly improve the framework as a description of the MME. Formally, if the p value of the test is small ($p < a$), we conclude that there is significant evidence of model dependence in the historical climate at the a% level and that the additive framework is most appropriate. Otherwise, the one-way framework is more appropriate.

h. Strength of evidence of climate change

When the expected climate response β_F is small, it may be difficult to distinguish it from the internal variability. In the appendix, a significance test is derived to test for the presence of a climate response signal, that is, to test the null hypothesis $\beta_F = 0$ against the alternative $\beta_F \neq 0$. The test statistic is

$$T_\beta = \frac{|\hat{\beta}_F|}{\sqrt{\mathrm{Var}(\hat{\beta}_F)}}. \tag{11}$$

If the p value of the test is small ($p < a$), we conclude that there is significant evidence of a nonzero climate response at the a% level of significance. If the p value is not small, we conclude that there is no significant evidence of a climate response.

The standardized effect size $d_\beta = |\hat{\beta}_F|/s$, where s is the estimate of σ, is a practical way of quantifying the size of the climate response. It is easily understood on the scale of the internal variability using the quantiles of the standard normal distribution, that is, $d_\beta \simeq 2$ implies the projected future climate is more extreme than 95% of plausible historical climates. The IPCC Fourth Assessment Report (Meehl et al. 2007a, Fig. 10.9) highlights climate responses greater than one standard deviation of intermodel spread. This is more closely related to f_γ^2 than to d_β, which is measured on the scale of internal variability. The value of d_β considered large for practical purposes may vary depending on the impact of a particular response. However, the scale is useful, and $d_\beta > 1$ represents a natural threshold for less impact focused studies.

i. Testing of individual climate models

Similar tests to that given for nonzero expected climate response in the previous section can be made for nonzero model dependence in the climate response ($\gamma_{mF} \neq 0$) and historical climate ($\alpha_m \neq 0$) of the individual models.

Under the null hypotheses of no model dependence in the historical climate of model m (H_0: $\alpha_m = 0$) and no model dependence in the climate response of model m (H_0: $\gamma_{mF} = 0$), the test statistics are

$$T_\alpha = \frac{|\hat{\alpha}_m|}{\sqrt{\mathrm{Var}(\hat{\alpha}_m)}} \quad \text{and} \quad T_\gamma = \frac{|\hat{\gamma}_{mF}|}{\sqrt{\mathrm{Var}(\hat{\gamma}_{mF})}}. \tag{12}$$

If the p value of one of these tests is small ($p < a$), we conclude that there is significant evidence at the a% level that model m differs from the ensemble mean in either its historical climate or climate response, that is, model m does not agree with the expected historical climate or climate response.

Removing models that disagree strongly with the expected climate or climate response runs the risk of not sampling unlikely yet still plausible climates. Therefore, models should not be excluded from the ensemble simply because they do not agree with the ensemble mean. It is useful to be able to systematically identify such models as such behavior may indicate problems that need to be investigated further. It may subsequently be decided that these problems warrant the exclusion of the model for the analysis of some or all climate variables. However, this should be based on expert judgment.

j. Framework selection strategy

The frameworks discussed in the previous sections form a hierarchy. The one-way framework is a special case of the additive framework, which is itself a special case of the two-way framework with interactions. A simple approach to selecting the most appropriate framework would be to calculate and compare the estimates of the expected climate response β_F from all three frameworks. The estimates may be obtained by simply calculating the weighted mean response in Eq. (2) using the weights in Eqs. (3), (6), and (8). If all three estimates are similar, then the one-way framework is probably sufficient to describe the MME. If the additive and two-way frameworks appear similar to each other but different to the one-way framework, then the additive framework is probably more appropriate. If all three frameworks give different estimates then either the two-way framework with interactions is required or a simple ANOVA framework is not appropriate.

A more rigorous approach would make use of the significance tests and assumption checking procedures outlined above:

1) Fit the two-way framework with interactions.
2) Check the framework assumptions and identify any outlying runs:
   (i) If the assumptions appear satisfied and there are no outlying runs, then go to the next step.
   (ii) If there are outlying runs, investigate possible causes before removing completely, or consider removing temporarily and rechecking the assumption of normality.
   (iii) If the assumptions do not appear satisfied and there are no outlying runs, then consider an alternative statistical framework or revert to the previous framework.
3) Perform the significance test for model dependence in the climate response. If the null hypothesis of no model dependence is rejected then stop; the two-way framework with interactions is most appropriate.
4) Fit the additive framework.
5) Check the framework assumptions and identify any outlying runs as in Step 2.
6) Perform the significance test for model dependence in the historical climate. If the null hypothesis of no model dependence is rejected then stop; the additive framework is most appropriate.
7) Fit the one-way framework.
8) Check the framework assumptions and identify any outlying runs as in Step 2.

Once the most appropriate framework has been selected, the test for nonzero climate response can be performed to identify whether or not there is significant evidence of a climate response in the MME. The values of d_β and f_γ^2 or f_α^2 may be examined in order to assess the size of the response and level of agreement between models.

Using the significance tests, the framework selection procedure may be easily automated for multiple grid points. Some manual intervention is required in checking the framework assumptions. The check for normality may be automated using the Anderson–Darling test. The Anderson–Darling test has greater power to detect a range of departures from normality than the more general Kolmogorov–Smirnoff test (Stephens 1974). The checks for constant variance and for independence should be performed at a random selection of grid points at each stage. When removing outliers, even temporarily, care must be taken to ensure that at least one run remains available under each scenario from each climate model.

3. Example: Storm tracks in CMIP5

a. Data

The frameworks outlined in the previous section are used to estimate changes in the 30-yr mean wintertime [December–February (DJF)] track density of

TABLE 2. Number of realizations available from each model for the historical and future scenarios and the weights given by each ANOVA framework. Weights have been standardized to sum to 100 for each framework.

	Runs		Weights					
			Two way		Additive		One way	
	Historical	RCP4.5						
Model	R_{mH}	R_{mF}	W_{mF}	W_{mH}	W_{nF}	W_{mH}	W_{mF}	W_{mH}
BCC-CSM1.1	3	1	2.63	2.63	2.25	2.25	3.85	1.28
CanESM2	5	1	2.63	2.63	2.50	2.50	6.41	1.28
CNRM-CM5	5	1	2.63	2.63	2.50	2.50	6.41	1.28
CSIRO-Mk3.6.0	4	5	2.63	2.63	6.68	6.68	5.13	6.41
EC-EARTH	3	3	2.63	2.63	4.51	4.51	3.85	3.85
FGOALS-g2	1	1	2.63	2.63	1.50	1.50	1.28	1.28
GFDL-ESM2G	1	1	2.63	2.63	1.50	1.50	1.28	1.28
GFDL-ESM2M	1	1	2.63	2.63	1.50	1.50	1.28	1.28
HadGEM2-CC	2	1	2.63	2.63	2.00	2.00	2.56	1.28
HadGEM2-ES	1	1	2.63	2.63	1.50	1.50	1.28	1.28
INM-CM4	1	1	2.63	2.63	1.50	1.50	1.28	1.28
IPSL-CM5A-LR	4	4	2.63	2.63	6.01	6.01	5.13	5.13
IPSL-CM5A-MR	1	1	2.63	2.63	1.50	1.50	1.28	1.28
MIROC5	1	1	2.63	2.63	1.50	1.50	1.28	1.28
MIROC-ESM	3	1	2.63	2.63	2.25	2.25	3.85	1.28
MIROC-ESM-CHEM	1	1	2.63	2.63	1.50	1.50	1.28	1.28
MPI-ESM-LR	3	3	2.63	2.63	4.51	4.51	3.85	3.85
MRI-CGCM3	5	1	2.63	2.63	2.50	2.50	6.41	1.28
NorESM1-M	3	1	2.63	2.63	2.25	2.25	3.85	1.28
Total	48	30	50.00	50.00	50.00	50.00	61.54	38.46

extratropical cyclones in the North Atlantic from an ensemble of climate models participating in the WCRP CMIP5 (Taylor et al. 2012). For a more complete discussion of climate change in the North Atlantic storm track in the CMIP5 MME, see Zappa et al. (2013b). Six-hourly output suitable for storm track analysis is available from 19 models from 12 centers. To maximize independence between models, it might be sensible to include only one model from each center (Rougier et al. 2012, manuscript submitted to *J. Amer. Stat. Assoc.*). However, Pennell and Reichler (2011) show that the effect of including same center models is limited, so in this example all models are included. Projections are compared from two 30-yr periods. The recent climate is represented by the mean of a 30-yr period from the historical experiment between December 1975 and February 2005. The future climate is analyzed conditionally on the Representative Concentration Pathway 4.5 (RCP4.5) midrange mitigation emissions scenario (Moss et al. 2010). The mean of a 30-yr period between December 2099 and February 2099 is analyzed. At least one realization is available from each model for each scenario. The total number of realizations available for each model-scenario pair is summarized in Table 2.

The analysis methodology is similar to that used in several previous studies of extratropical cyclones (e.g., Bengtsson et al. 2006, 2009; Catto et al. 2011; McDonald 2011). Cyclones are identified as maxima in the 850-hPa relative vorticity field and tracked through their life cycle using the method developed by Hodges (1994, 1995, 1999). Prior to tracking, the large-scale background field is removed (Hoskins and Hodges 2002; Anderson et al. 2003). The output of the models is also interpolated to a common resolution of T42. This simplifies comparison between models and reduces the noise in the vorticity field. After tracking, storms that last less than 2 days or travel less than 1000 km are excluded. Spatial statistics are then computed from the tracks using the spherical kernel approach of Hodges (1996).

This example focuses on the track density statistics. This is the mean number of cyclones passing a particular point each month. The spherical kernel approach utilizes a variable bandwidth so the statistics are rescaled to be representative of a region of radius 5° centered on a particular grid point. This study focuses on the DJF winter period in the North Atlantic. The study region is defined as 80°E–40°W and 30°–90°N. This window covers the North Atlantic storm track and its exit region over Europe.

b. *Results*

1) THE SIMPLE APPROACH TO FRAMEWORK SELECTION

The simple approach to framework selection is illustrated in Fig. 1. The CMIP5 models simulate the DJF storm track reasonably well, but with some departures.

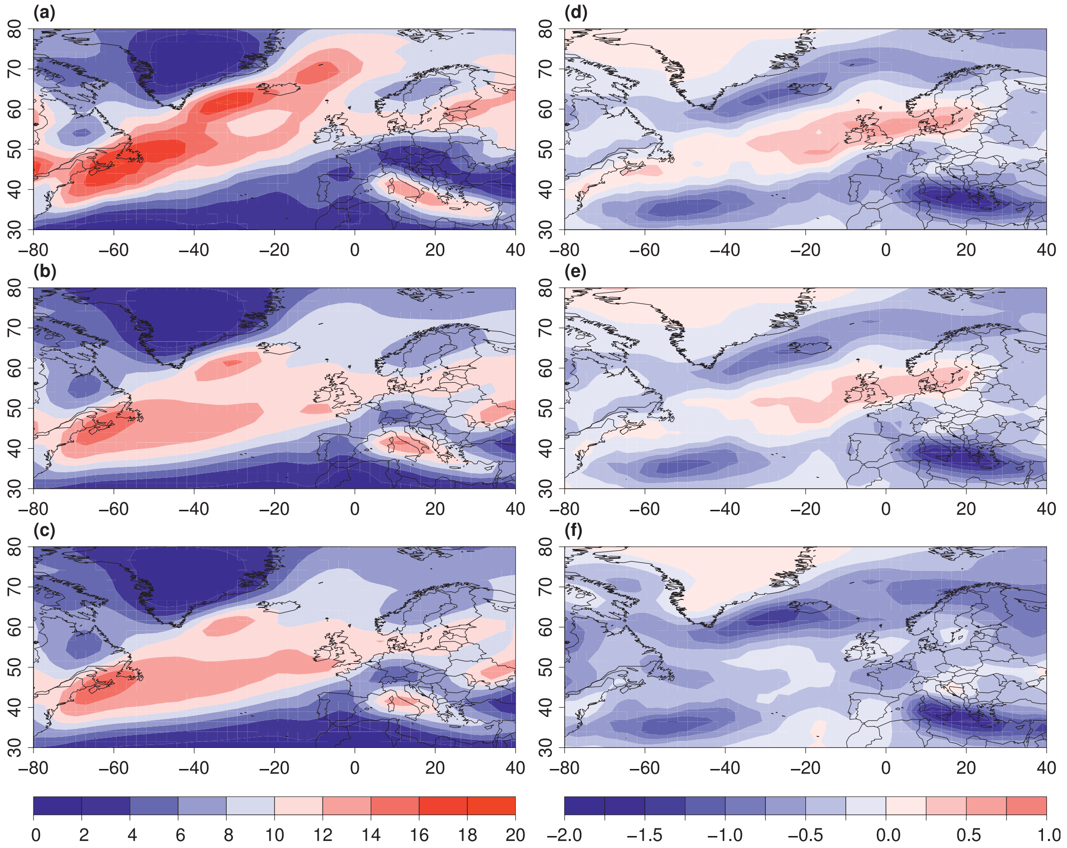

FIG. 1. (a) DJF track density (storms month^{-1}) in the European Centre for Medium-Range Weather Forecasts (ECMWF) Interim Re-Analysis (ERA-Interim), (b) CMIP5 expected historical DJF track density estimate from the two-way framework with interactions, and (c) CMIP5 expected RCP4.5 DJF track density estimate from the two-way framework with interactions. Expected climate response estimates (storms month^{-1}) (d) from the two-way framework with interactions, (e) from the additive framework, and (f) from the one-way framework.

The main northeast track is too weak, while the more zonal track toward northern Europe is too strong. Comparing the climate response estimates from the three frameworks in Figs. 1d–f suggests the additive framework may be suitable to describe the CMIP5 MME. The response estimates from the two-way framework with interactions and the additive framework appear similar. The response estimate from the one-way framework fails to capture the increase in track density over the United Kingdom and Denmark indicated by the other two frameworks. This suggests the presence of differences between the historical climates simulated by the CMIP5 models.

2) A SINGLE GRID POINT

To better understand the differences between the ANOVA frameworks, a single grid point in central France (46.5°N, 1.25°E) is considered in detail. Figure 2 confirms that there are large differences between the historical climates simulated by the CMIP5 models. By comparison, the usual one model, one vote estimate of the climate response indicated by the horizontal dashed lines is small. Where multiple runs are available, the spread appears comparable between models and scenarios. This suggests the assumption of constant variance is justified for cyclone track density in the CMIP5 MME. One exception is the MIROC-ESM model, which appears to have an unusually large spread of values in the historical scenario at this grid point. Most models appear to show a small decrease in track density in the RCP4.5 scenario compared to the historical scenario. However, there is some variation in the size of the decrease. The two-way framework with interactions may be required to explain this variation if it is greater than might be expected because of internal variability.

The differences between the structures of the three ANOVA frameworks are also visible in Fig. 2. In the two-way framework with interactions, the ML estimate of the mean climate in each model and scenario is the sample mean of the runs from that model-scenario pair

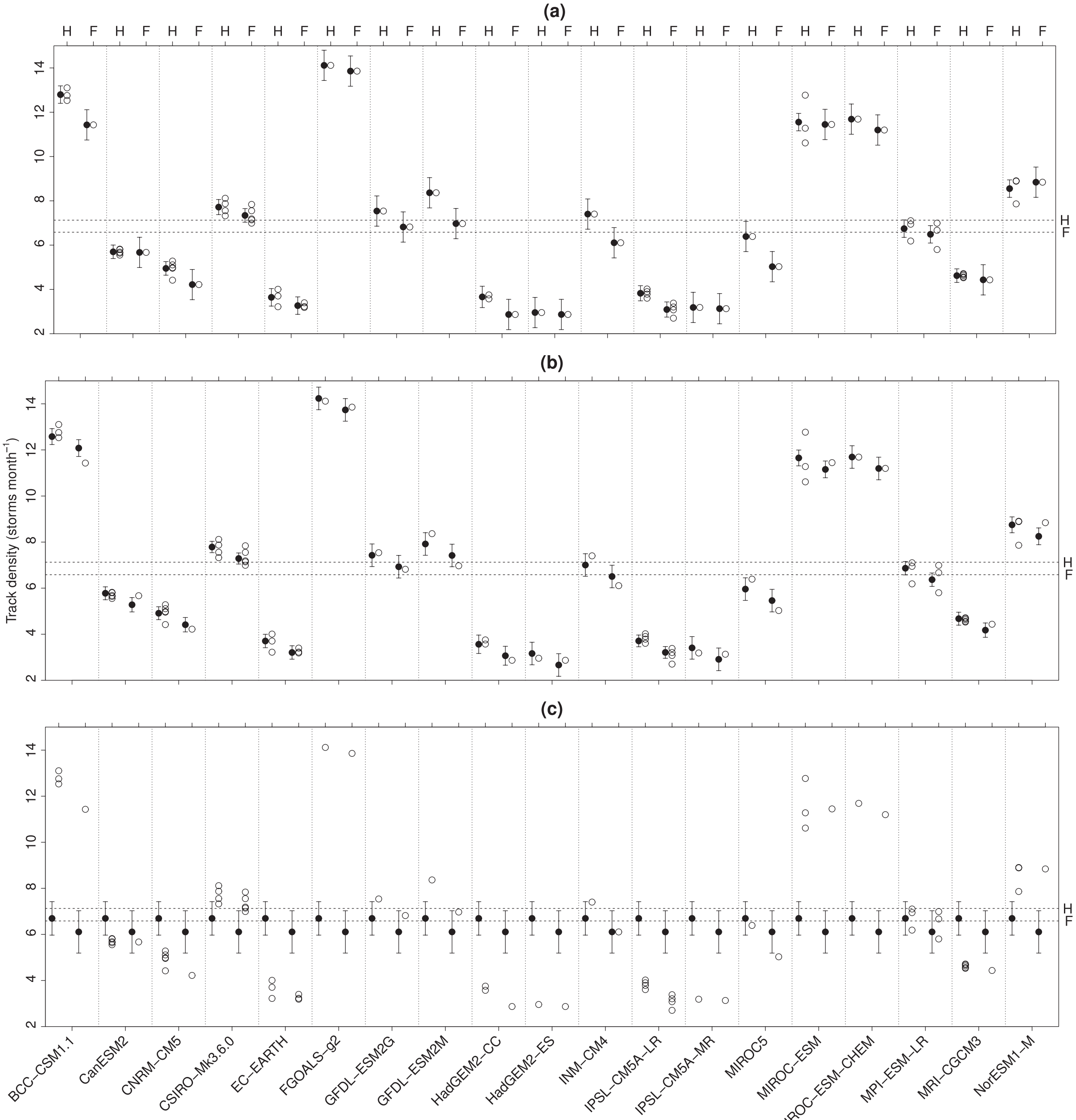

FIG. 2. Estimated mean climates from the three ANOVA frameworks for a grid point (46.5°N, 1.25°E) in central France: track density vs model. (a) The two-way framework with interactions, (b) the additive framework, and (c) the one-way framework. Open points represent individual runs from the historical scenario (*H*, left in each column) and the RCP4.5 (future) scenario (*F*, right) for each model. Solid points are framework estimates of the mean climate of each model for each scenario. Error bars represent a 90% confidence interval for the mean climate of each model. Dashed horizontal lines indicate the usual one model, one vote estimates of the historical and future climates.

(Fig. 2a). Different climate responses are estimated for each model. The additive framework constrains the estimates so that all models have the same climate response. While no longer centered on the model-scenario means, these estimates appear reasonable for most models (Fig. 2b). The uncertainty indicated by the error bars is reduced compared to the two-way framework with interactions, suggesting that the additive framework may be sufficient to describe the MME. The one-way framework constrains the estimates so that all models simulate the same historical climate and climate response (Fig. 2c). Note that these do not coincide with the usual one model, one vote estimates. The error bars indicate that the uncertainty is greater than in the additive framework. This is not surprising given the large differences between the historical climates simulated by the CMIP5 models. These are not captured at all by the one-way framework and are therefore absorbed into the estimate of the internal variability.

The expected climate response estimate $\hat{\beta}_F$ and associated 90% confidence interval from the two-way framework with interactions is -0.54 (-0.73, -0.36) storms month^{-1}. From the additive framework the estimate $\hat{\beta}_F$ is -0.50 (-0.67, -0.33) storms month^{-1}, and from the one-way framework it is -0.59 (-1.76, 0.58) storms month^{-1}. The decrease in width of the confidence intervals in the additive framework suggests that the interaction terms are not required in order to adequately describe the uncertainty in the ensemble. The dramatic increase in width of the confidence intervals from the one-way framework reflects the large structural uncertainty in historical climate that has been absorbed into the estimate of the internal variability.

No systematic patterns are visible in the plot of standardized residuals against the fitted values from the two-way framework with interactions in Fig. 3a. This suggests the assumption of constant variance is justified. Two outlying runs are indicated from the MIROC-ESM model. The same runs are indicated in the quantile–quantile plot in Fig. 3b. Most runs lie close to the expected straight line, although some skewness is visible. This is likely to be because of the influence of the two outliers. After removing the two runs of MIROC-ESM, no further outliers are identified. The p value of the Anderson–Darling test for normality is 0.16, so there is no significant evidence to reject the null hypothesis of normality. Investigating the reasons behind the two outlying runs of MIROC-ESM is beyond the scope of this example. Removing the two outliers has very little effect on the estimates of the main effects μ and β. We therefore proceed with the two outlying runs included in the ensemble, but we are reassured that the framework assumptions are basically justified at this grid point.

The variance ratio f_γ^2 is calculated as 0.47, that is, structural uncertainty in the climate response explains variability equivalent to 47% of that explained by internal variability. The p value of the significance test for model-dependent climate response based on f_γ^2 is 0.44. There is no evidence to reject the null hypothesis of no model-dependent climate response at the 10% level. Therefore, the additive framework may be adequate to describe the variability in the MME. Checking the framework assumptions under the additive framework reveals no problems. The variance ratio f_α^2 is calculated as 70.5, that is, structural uncertainty in the historical climate explains 71 times more variation in the CMIP5 MME than the internal variability. This result is highly significant. The null hypothesis of no model dependence in the historical climate is rejected entirely. At this grid point, the additive framework provides the most parsimonious description of the MME.

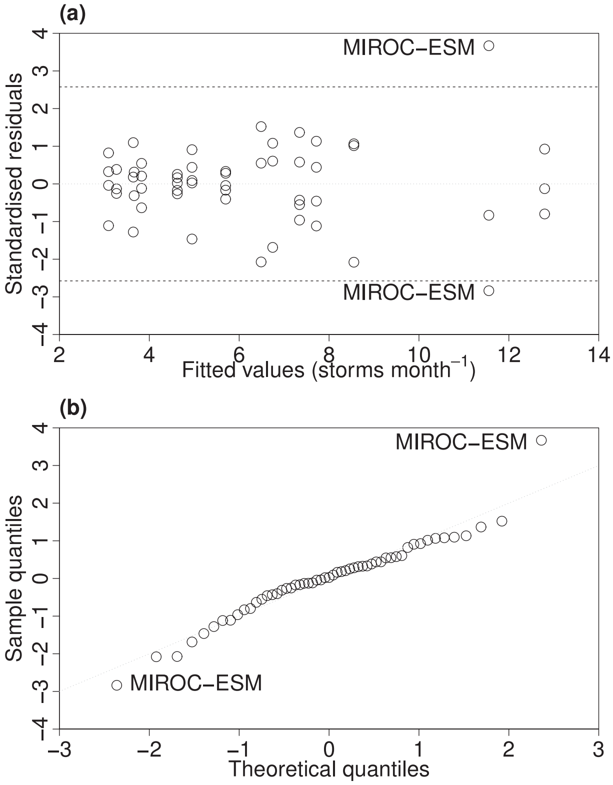

FIG. 3. Framework checking for the two-way framework with interactions. (a) Plot of standardized residuals against fitted values. Each point represents one run. Dashed lines indicate the 0.5% and 99.5% quantiles of the standard normal distribution. (b) Quantile–quantile plot of the standardized residuals.

3) THE NORTH ATLANTIC STORM TRACK

Figure 1 suggests that the structural uncertainty in the climate response of the CMIP5 ensemble is small despite large structural uncertainty in the historical climate; that is, the models agree on the climate response but not the historical climate. Before the hypothesis of no model-dependent climate response can be tested, the framework assumptions must be checked in the two-way framework with interactions.

Plots of standardized residuals against fitted values at a random selection of grid points (not shown) reveal no

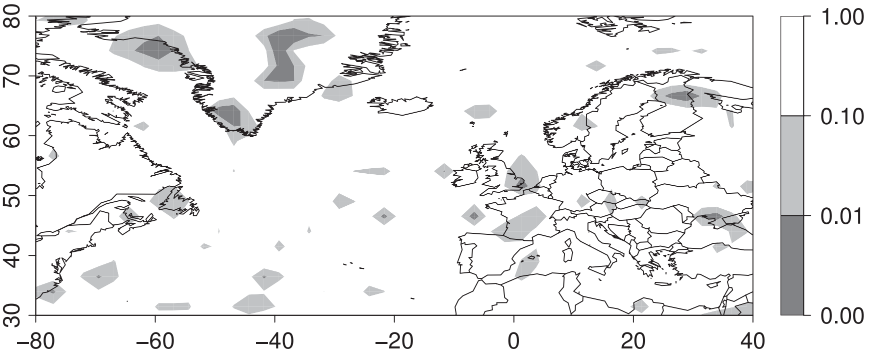

FIG. 4. The p values of the Anderson–Darling test for normality in the two-way framework with interactions. Small p values ($p < 0.10$) indicate significant evidence of nonnormality. The assumption of normality appears justified over most of the study region.

evidence of nonconstant variance between models or scenarios. The Anderson–Darling test (Fig. 4) suggests that the assumption of normality is justified over most of the study region. The two outlying runs identified over central France do not persist across the study region. To perform a thorough check for outlying runs, the standardized residuals of each run were mapped individually ($N = 78$ plots, not shown). No single run is identified as outlying at the 1% level at more than 4% of grid points, and these are usually spread over multiple subregions. Therefore, we proceed with all runs included in the ensemble.

The variance ratio f_γ^2 and p values of the significance test for model-dependent climate response are shown in Figs. 5a and 5b. The structural uncertainty associated with the climate response is less than the uncertainty due to internal variability over most of the study region. However, areas of significant nonzero model dependence at the 10% level are detected, most notably over the subtropical Atlantic Ocean, away from the main storm track.

To determine which models are not in agreement with the rest of the CMIP5 MME, the outcomes of the significance tests on the individual γ_{mF} effects [Eq. (12)] are mapped in Fig. 6. No one model or group of models appears responsible for all of the interaction in the climate response. Different groups of models deviate from the rest of the MME in different regions. In the subtropical Atlantic Ocean, CSIRO-Mk3.6.0, FGOALS-g2, MIROC-ESM, and MIROC-ESM-CHEM all deviate strongly from the expected response. MRI-CGCM3 is unique in that it deviates from the expected response near the Iberian Peninsula, but not over the rest of the subtropical Atlantic.

Figure 6 indicates that all the regions of interaction detected in Fig. 5b involve more than one model. Comparing plots in Fig. 6 shows that models that share similar responses in one area will not necessarily have similar responses in another. Therefore, removing any model from the MME entirely would remove useful information in some regions and risk excluding unlikely but still plausible climate responses in other regions.

Although there is evidence of structural uncertainty in the climate response in some areas, there is good agreement between models over most of the study region. Where the structural uncertainty is small compared to the internal variability, the additive framework may provide a more parsimonious description of the MME. Fitting the additive framework and checking the assumptions (not shown) reveals no problems.

However, examining the variance ratio f_α^2 (not shown) reveals that even where the models agree on the climate response, there are large differences in their historical climates. Differences among the historical climates of the models are responsible for at least twice the variation explained by the internal variability everywhere in the study region. Over central Europe the variance ratio rises to $f_\alpha^2 \approx 70$. This agrees with Zappa et al. (2013a), who found that the storm tracks of several models extend too far into the European continent. On the basis of this evidence, the one-way framework, where runs are weighted equally, should not be used to estimate the climate response anywhere in the North Atlantic storm track.

The difference between the estimates of the expected climate response β_F from the two-way framework with interactions and the additive framework is shown in Fig. 7a. A comparison with Figs. 1a and 1b shows that the two-way framework with interactions tends to estimate a stronger climate response than the additive framework. Since both estimates are weighted averages, the difference must be due to the weights. In Table 2 the additive framework assigns most weight to the CSIRO-Mk3.6.0, EC-EARTH, IPSL-CM5A-LR,

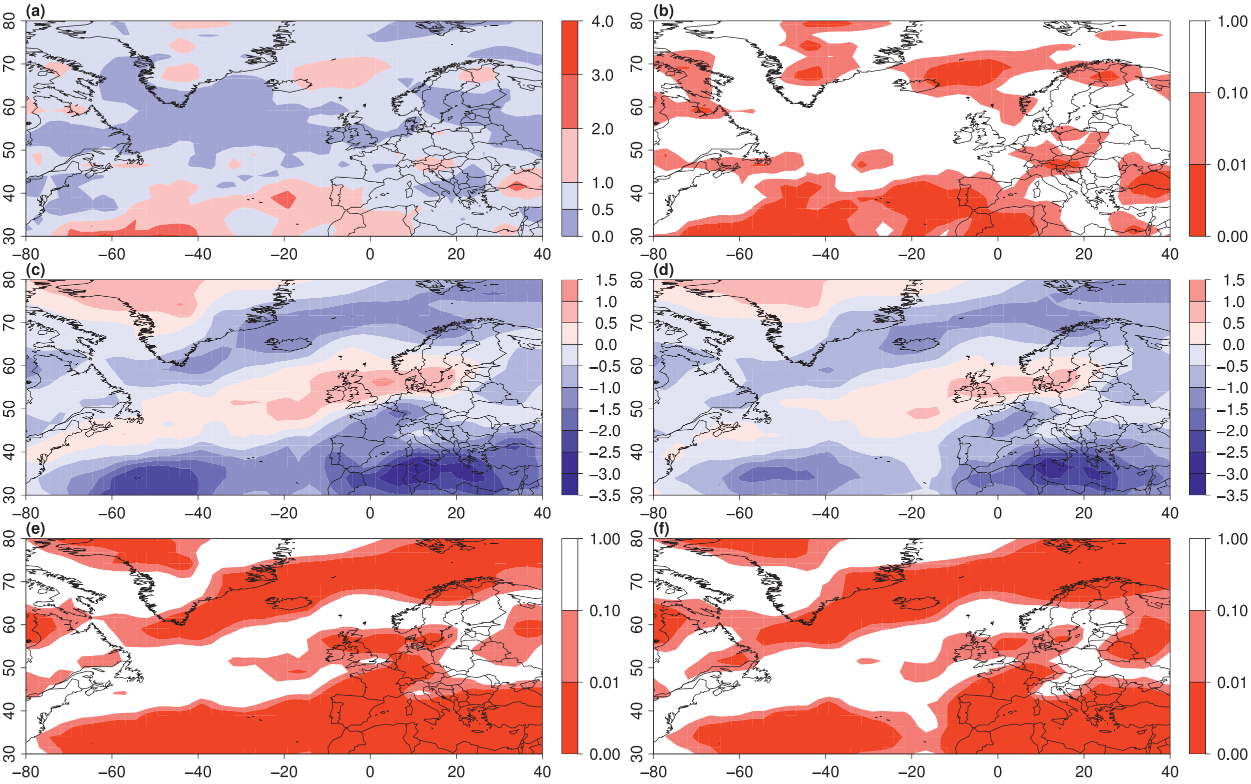

FIG. 5. (a) Variance ratio f_γ^2, (b) p values of the significance test for model-dependent climate response, (c) standardized mean climate response d_β from the two-way framework, (d) standardized mean climate response d_β from the additive framework, (e) p value of significance test for nonzero mean climate response from the two-way framework, and (f) p value of the significance test for nonzero mean climate response from the additive framework.

and MPI-ESM-LR models. Comparing these models in Fig. 2a shows that they all have relatively weak climate responses (at this grid point). Since the additive framework gives increased weight to these models, its climate response estimate is correspondingly lower.

Figure 7b shows the spatial distribution of the standard error of the expected climate response estimate $\hat{\beta}_F$ from the additive model. Variability decreases away from the main storm track. This is to be expected since the standard errors from the ANOVA frameworks represent the uncertainty due to internal variability. A 90% confidence interval for the expected climate response β would have width ±1.64 SE, where SE is the standard error from Fig. 7b. Local maxima in the standard error over Newfoundland, Denmark, Corsica, and near the tip of Greenland appear related to areas of strong primary or secondary cyclogenesis (Hodges et al. 2011).

Comparing Fig. 7c with Fig. 5b shows that the expected climate response estimate from the additive framework has greater precision than the two-way framework with interactions where there is no significant evidence of model dependence in the response. Note that the decrease in precision from using the two-way framework with interactions where there is no evidence of model dependence in the response is generally small compared to the decrease in precision from using the additive framework where there is model dependence. This agrees with the theoretical arguments in section 2c.

Both the two-way and additive frameworks estimate large ($d_\beta > 1$) climate responses in the subtropical North Atlantic, the Mediterranean, and parts of the main northeast branch of the storm track (Figs. 5c,d). The statistical significance of these responses is shown in Figs. 5e and 5f. Both frameworks find significant evidence of nonzero climate response at the 1% level over the three regions already highlighted plus France, Spain, Portugal, Switzerland, and parts of northern Europe.

In the CMIP5 MME, there is significant evidence of a decrease in the frequency of cyclones on the northern flank of the North Atlantic storm track in the RCP4.5 scenario. A significant increase is also noted on the southern flank. However, there is significant structural uncertainty in the climate response in this region, so the result should be treated with caution. A small increase in frequency is indicated in the zonal branch of the storm

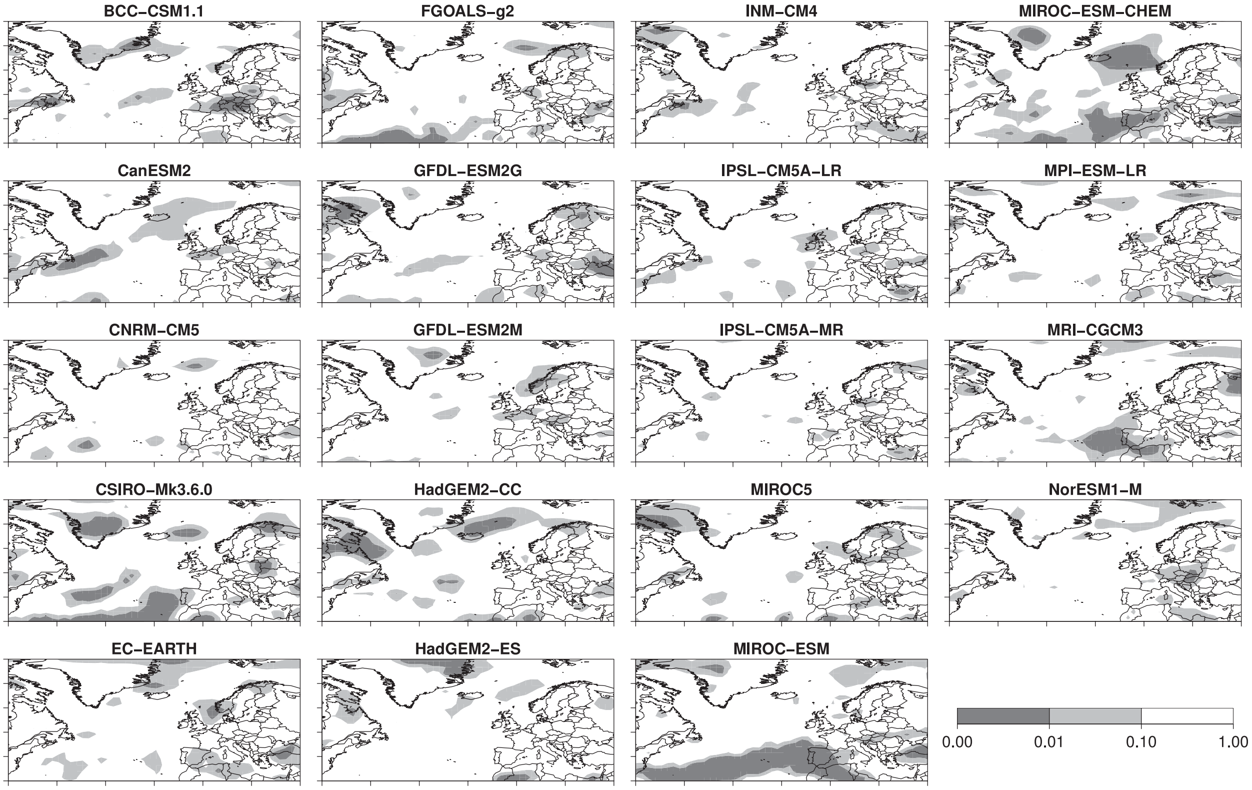

FIG. 6. The p values of the individual t tests on the γ_{mF} terms. Small p values ($p < 0.10$) indicate significant evidence that a particular model disagrees with the mean climate response of the MME.

track directed toward northern Europe. There is evidence at the 10% level, but not at the 1% level, that this increase is because of a change in radiative forcing rather than internal variability. However, the evidence is not strong enough that we can be certain. The largest responses are seen in the Mediterranean basin. In this region a decrease in storm frequency of up to two storms per month is projected. This corresponds to a standardized decrease of up to three standard deviations, a very strong signal. This could have serious consequences for water supplies in southern Europe and the Middle East.

4. Conclusions

This study describes a family of ANOVA frameworks, the most general of which naturally yields the one model, one vote estimate of future climate response in a MME. Two alternative estimates, including a one run, one vote estimate, are also introduced, and they are more efficient when the structural uncertainty is small compared to the internal variability. The assumptions of these frameworks can be rigorously checked using simple tests and graphical techniques. The ANOVA frameworks allow the construction of confidence intervals in addition to the usual point estimates. The frameworks described here overcome the need to analyze only one run, or an equal number of runs, from each model-scenario pair by using linear regression techniques rather than traditional ANOVA estimation.

The two-way ANOVA framework with interactions shows that the one model, one vote estimate of the ensemble mean climate response implicitly allows for the possibility that each climate model may respond differently to the same radiative forcing. If the models all respond differently, it is difficult justify the ensemble mean climate response as an estimate of the actual climate response without assuming that the models are truth centered. However, this assumption is often hard to justify (Knutti et al. 2010b).

The principle behind the use of MMEs for climate projection is that each model represents a line of evidence for the future behavior of the actual climate. If multiple lines of high-quality evidence agree, then confidence is increased (Mastrandrea et al. 2010). Therefore, it is hoped that a consensus will exist among climate models on the climate change response. If the models all simulate the same climate response, no truth-centered assumption is required in order to justify that

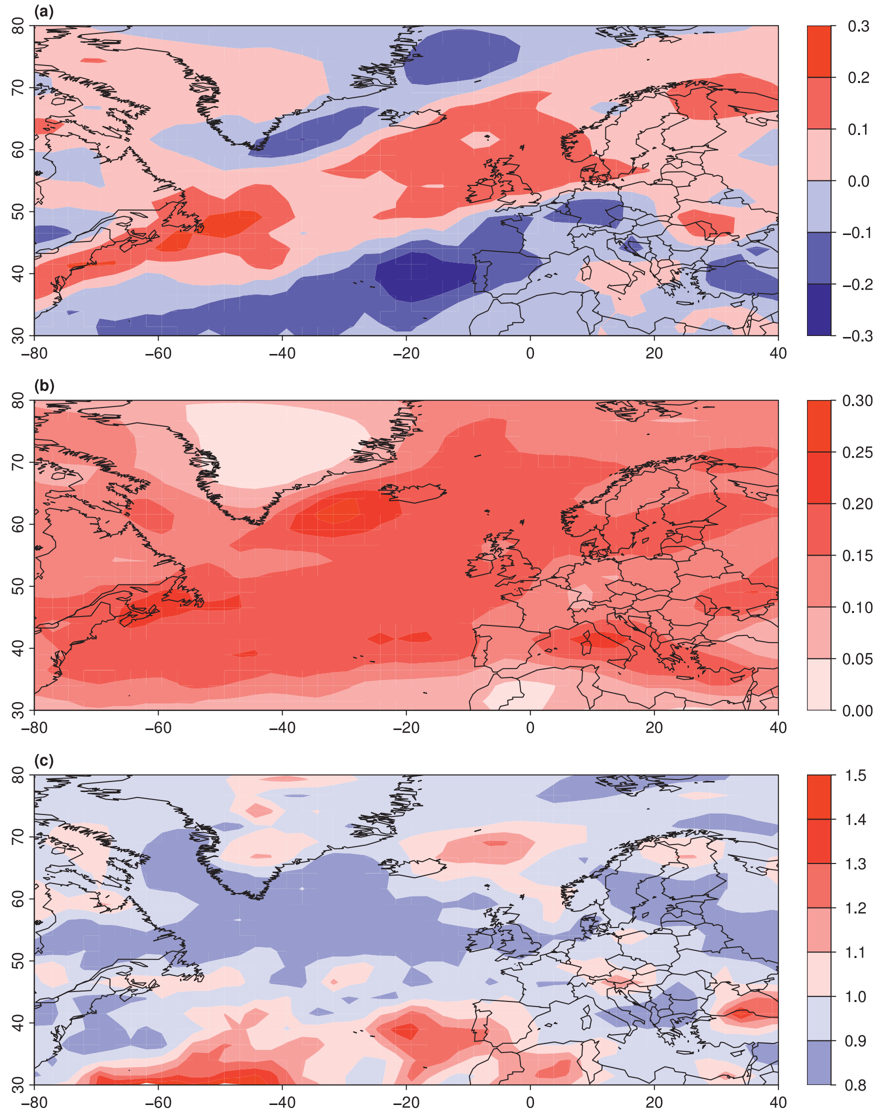

FIG. 7. (a) Difference between the expected climate response estimates (storms month^{-1}) from the two-way framework with interactions and the additive framework, (b) standard error (storms month^{-1}) of expected climate response estimate from the additive framework, and (c) ratio of standard errors of the expected climate response estimates from the additive framework and the two-way framework with interactions.

response as an estimate of the actual climate response. However, it is necessary to assume that any bias shared by all the models compared to the actual climate is constant between historical and future scenarios. Such an assumption is more acceptable than the increasingly unsupportable truth-centered approach (Stephenson et al. 2012).

The additive ANOVA framework assumes that all the models simulate the same climate response, even if they simulate different historical mean climates. When this assumption is justified, the associated estimate of the climate response will have greater precision than the one model, one vote estimate, as well as being more defensible as an estimate of the actual climate response. The ML estimate of the climate response from this framework is a weighted average of the sample mean responses from the individual models. The model weights depend on the number of runs from each model-scenario pair. Having many runs from only one scenario does not yield a high weighting. This emphasizes the need for modeling centers to provide multiple runs from future scenarios, not just the historical scenario.

This study shows that the assumption that all models in the MME simulate the same climate response can be formally justified based on the ratio of the structural uncertainty in the climate response to the uncertainty due to internal variability. When the ratio is small, there is insufficient evidence (i.e., runs) in the MME to reliably distinguish any structural uncertainty in the response from internal variability. It is unlikely that models will ever agree completely on the climate response; however, it can be hoped that the differences are small. Given sufficient runs, even small differences could be distinguished. However, if they are sufficiently small compared to the internal variability, then their estimation may be safely neglected.

There is increasing awareness of the role of internal variability in climate projection (Deser et al. 2012; Tebaldi et al. 2011a; Yip et al. 2011). In agreement with Tebaldi et al. (2011a), we find that the agreement between models on the future climate response may be greater than previously thought. In particular, the methods presented here argue strongly against the practice of selecting only one run, or a subset of runs, from each model-scenario pair when additional runs are available.

The example of the North Atlantic storm track demonstrates that, to within the range of internal variability, climate models generally agree on the extratropical cyclone frequency response to the RCP4.5 scenario in DJF. This is surprising considering that climate models often simulate different storm tracks in the historical scenario (Zappa et al. 2013a). We also demonstrate how outlying runs and models may be systematically identified for further investigation. However, such runs or models should only be removed from the ensemble based on expert judgment following a detailed investigation of the underlying cause of the discrepancy.

When applying the significance tests on a grid point basis, as in section 3, it may be necessary to consider spatial dependence in the data. We do not address this explicitly here. However, depending on the application, authors may wish to consider applying either the field significance method (Livezey and Chen 1983) or the false discovery rate method (Ventura et al. 2004) to account for any dependence.

ANOVA frameworks can be used to quantify the relative contributions of the various components of uncertainty to the total uncertainty in an MME (Yip et al. 2011). However, only the internal variability is quantified absolutely. If the structural uncertainty in the climate response is small compared to the internal variability, confidence intervals for the climate response from the additive framework should be reported. This should be accompanied by a statement of the assumption of constant shared bias. However, subject to that assumption, we should have high confidence in such an estimate.

When the models do not agree on the climate response, only the usual one model, one vote estimate of the climate response should be reported. This should be accompanied by a statement of limited confidence in the findings. To report confidence intervals from the two-way framework with interactions would be to ignore the structural uncertainty in the climate response as well as the uncertainty because of any biases shared by all the models. This would give an impression of false confidence. The method of Tebaldi et al. (2011a) could be employed in conjunction with the significance tests in section 2i to visualize the level of agreement between models.

When the agreement on the climate response is poor, the challenge is to determine the scientific reasons for the differences between the models. In some cases feedbacks exist that cause the climate response simulated by a particular model to depend strongly on the historical climate in that model (e.g., Bracegirdle and Stephenson 2012). In other cases, particular variables might be found to depend strongly on other processes that vary between models (e.g., Woollings et al. 2012). If such relationships have a physical basis, they could be incorporated into the statistical framework in order to reduce the structural uncertainty.

Ideally, we would like to make quantitative statements about the uncertainty in the climate response, even in the presence of shared biases and when models do not agree on the response. However, to do so would require a subjective view of the nature of probability in order to express the size of the structural uncertainty. It is difficult to imagine a notional population of climate models from which the models in the ensemble were sampled (Stephenson et al. 2012), so the ideas of classical statistics do not apply. A number of Bayesian hierarchical frameworks (Chandler 2013; Rougier et al. 2012, manuscript submitted to *J. Amer. Stat. Assoc.*; Tebaldi et al. 2011b) have been proposed that allow this subjective evaluation. However, this study has shown that such complex frameworks are not always necessary in order to quantify the uncertainty in climate change projections.

At present, there is no consensus on a ''correct'' framework for quantifying the uncertainty in climate projections from MMEs. Both the simple ANOVA frameworks outlined here and the more complex Bayesian frameworks suggested elsewhere make assumptions about the independence of models. The issue of how the biases between models and the actual climate may evolve in the future is also an area of active research (Stephenson et al. 2012). Identifying outlying runs and models and the incorporation of physical relationships into statistical frameworks all point to the importance of process-based evaluation of climate models. Process-based comparisons also suggest an alternative approach to model weighting. Incorporating such process-based information will only be achieved by increased cooperation between statisticians and climate scientists.

Acknowledgments. This study was conducted as part of a studentship funded by the Natural Environment Research Council under the Testing and Evaluating Model Predictions of European Storms (TEMPEST) project. We acknowledge the World Climate Research Programme's Working Group on Coupled Modelling, which is responsible for CMIP, and we thank the climate modeling groups (listed in Table 1 of this paper) for producing and making available their model output. For CMIP the U.S. Department of Energy's Program for Climate Model Diagnosis and Intercomparison provides coordinating support and led development of software infrastructure in partnership with the Global Organization for Earth System Science Portals.

APPENDIX

Estimates, Standard Errors, Significance Tests, and Confidence Intervals

a. Derivation of two-way framework with interactions

The log likelihood of the two-way framework with interactions in Eq. (4) is

$$l(\mu, \alpha_m, \beta_s, \gamma_{ms}, \sigma^2; \mathbf{y}) = -\frac{N}{2}\log(2\pi) - N\log(\sigma) - \frac{1}{2\sigma^2}\sum_{m=1}^{M}\sum_{s\in\{H,F\}}\sum_{r=1}^{R_{ms}}(y_{msr} - \mu - \alpha_m - \beta_s - \gamma_{ms})^2, \tag{A1}$$

with the usual constraints $\sum_{m=1}^{M}\alpha_m = 0$, $\beta_H = \gamma_{mH} = 0$ $\forall\, m$, and $\sum_{m=1}^{M}\gamma_{mF} = 0$. ML estimates are obtained by maximizing the log likelihood with respect to all the parameters simultaneously. This is equivalent to solving the set of simultaneous equations arising from partial differentiation of the log likelihood with respect to each parameter and setting each equation equal to zero. Solving the set of simultaneous equations yields the following estimates:

$$\hat{\mu} = \frac{1}{M}\sum_{m=1}^{M}\overline{y}_{mH.}, \tag{A2a}$$

$$\hat{\alpha}_m = \overline{y}_{mH.} - \hat{\mu}, \tag{A2b}$$

$$\hat{\beta}_F = \frac{1}{M}\sum_{m=1}^{M}(\overline{y}_{mF.} - \overline{y}_{mH.}), \tag{A2c}$$

$$\hat{\gamma}_{mF} = (\overline{y}_{mF.} - \overline{y}_{mH.}) - \hat{\beta}_F, \tag{A2d}$$

and

$$s^2 = \hat{\sigma}^2 = \frac{1}{N-P}\sum_{m=1}^{M}\sum_{s\in\{H,F\}}\sum_{r=1}^{R_{ms}}(y_{msr} - \hat{y}_{msr})^2, \tag{A3}$$

where $P = 2M$ is the number of effects to be estimated and $\hat{y}_{msr} = \hat{\mu} + \hat{\alpha}_m + \hat{\beta}_s + \hat{\gamma}_{ms}$. The variances of the estimates are given by

$$\mathrm{Var}(\hat{\mu}) = \frac{\sigma^2}{M^2}\sum_{m=1}^{M}\frac{1}{R_{mH}}, \tag{A4a}$$

$$\mathrm{Var}(\hat{\alpha}_m) = \mathrm{Var}(\hat{\mu}) + \frac{\sigma^2}{R_{mH}}\left(\frac{M-2}{M}\right), \tag{A4b}$$

$$\mathrm{Var}(\hat{\beta}_F) = \frac{\sigma^2}{M^2}\sum_{m=1}^{M}\left(\frac{R_{m.}}{R_{mH}R_{mF}}\right), \tag{A4c}$$

$$\mathrm{Var}(\hat{\gamma}_{mF}) = \mathrm{Var}(\hat{\beta}_F) + \sigma^2\left(\frac{R_{m.}}{R_{mH}R_{mF}}\right)\left(\frac{M-2}{M}\right), \tag{A4d}$$

and

$$\mathrm{Var}(\hat{y}_{msr}) = \sigma^2/R_{ms}, \tag{A4e}$$

where $R_{m.} = R_{mH} + R_{mF}$. However, σ^2 is unknown, so it is replaced by the estimate s^2 from Eq. (A3).

b. Derivation of additive framework

The log likelihood of the additive framework in Eq. (5) is

$$l(\mu, \alpha_m, \beta_s, \sigma^2; \mathbf{y}) = -\frac{N}{2}\log(2\pi) - N\log(\sigma) - \frac{1}{2\sigma^2}\sum_{m=1}^{M}\sum_{s\in\{H,F\}}\sum_{r=1}^{R_{ms}}(y_{msr} - \mu - \alpha_m - \beta_s)^2, \tag{A5}$$

with the usual constraints $\sum_{m=1}^{M}\alpha_m = 0$ and $\beta_H = 0$. Estimation proceeds as for the two-way framework with interactions. Solving the set of simultaneous equations yields the ML estimates

$$\hat{\mu} = \frac{1}{M}\sum_{m=1}^{M}\left(\overline{y}_{m..} - \frac{R_{mF}}{R_{m.}}\hat{\beta}_F\right), \tag{A6a}$$

$$\hat{\alpha}_m = \left(\overline{y}_{m..} - \frac{R_{mF}}{R_{m.}}\hat{\beta}_F\right) - \hat{\mu}, \tag{A6b}$$

and

$$\hat{\beta}_F = \frac{1}{\sum_{m=1}^{M} W_m} \sum_{m=1}^{M} W_m(\overline{y}_{mF.} - \overline{y}_{mH.}), \qquad \text{(A6c)}$$

where

$$W_m = \frac{R_{mH}R_{mF}}{R_{mH} + R_{mF}}. \qquad \text{(A7)}$$

The variances of the estimates are given by

$$\mathrm{Var}(\hat{\mu}) = \frac{\sigma^2}{M^2}\left[\sum_{m=1}^{M}\frac{1}{R_{m.}} + \frac{1}{\sum_{m=1}^{M} W_m}\left(\sum_{m=1}^{M}\frac{R_{mF}}{R_{m.}}\right)^2\right], \qquad \text{(A8a)}$$

$$\mathrm{Var}(\hat{\alpha}_m) = \frac{\sigma^2}{\sum_{m=1}^{M} W_m}\left(\frac{R_{mF}}{R_{m.}} - \frac{1}{M}\sum_{m=1}^{M}\frac{R_{mF}}{R_{m.}}\right)^2 + \frac{\sigma^2}{M^2}\sum_{m=1}^{M}\frac{1}{R_{m.}} + \frac{\sigma^2}{R_{m.}}\left(\frac{M-2}{M}\right), \qquad \text{(A8b)}$$

$$\mathrm{Var}(\hat{\beta}_F) = \frac{\sigma^2}{\sum_{m=1}^{M} W_m}, \qquad \text{(A8c)}$$

and

$$\mathrm{Var}(\hat{y}_{msr}) = \frac{\sigma^2}{R_{m.}} + \frac{\sigma^2}{R_{m.}^2 \sum_{m=1}^{M} W_m}(R_{m.}^2 - R_{ms}^2 - 2R_{m.}W_m), \qquad \text{(A8d)}$$

but σ^2 is unknown, so it is replaced by the estimate s^2 from Eq. (A3) with $P = M + 1$ and $\hat{y}_{msr} = \hat{\mu} + \hat{\alpha}_m + \hat{\beta}_s$.

c. Derivation of one-way framework

The log likelihood of the one-way framework in Eq. (7) is

$$l(\mu, \alpha_m, \sigma^2; \mathbf{y}) = -\frac{N}{2}\log(2\pi) - N\log(\sigma) - \frac{1}{2\sigma^2}\sum_{m=1}^{M}\sum_{s\in\{H,F\}}\sum_{r=1}^{R_{ms}}(y_{msr} - \mu - \beta_s)^2, \qquad \text{(A9)}$$

with the usual constraint $\beta_H = 0$. Estimation proceeds as for the two-way framework with interactions. Solving the set of simultaneous equations yields the ML estimates

$$\hat{\mu} = \frac{1}{\sum_{m=1}^{M} R_{mH}}\sum_{m=1}^{M} R_{mH}\overline{y}_{mH.}, \quad \text{and} \qquad \text{(A10a)}$$

$$\hat{\beta}_F = \frac{1}{\sum_{m=1}^{M} R_{mF}}\sum_{m=1}^{M} R_{mF}\overline{y}_{mF.} - \hat{\mu}. \qquad \text{(A10b)}$$

The variances of the estimates are given by

$$\mathrm{Var}(\hat{\mu}) = \frac{\sigma^2}{\sum_{m=1}^{M} R_{mH}}, \qquad \text{(A11a)}$$

$$\mathrm{Var}(\hat{\beta}_F) = \frac{\sigma^2}{\sum_{m=1}^{M} R_{mH}} + \frac{\sigma^2}{\sum_{m=1}^{M} R_{mF}}, \qquad \text{(A11b)}$$

and

$$\mathrm{Var}(\hat{y}_{msr}) = \sigma^2/R_{.s}, \qquad \text{(A11c)}$$

but σ^2 is unknown, so it is replaced by the estimate s^2 from Eq. (15) with $P = 2$ and $\hat{y}_{msr} = \hat{\mu} + \hat{\beta}_s$.

d. The F tests for model dependence in the historical climate and climate response of the ensemble

The standard theory of the normal linear model (Krzanowski 1998) states that F_γ has a F distribution with $M - 1$ and $N - 2M$ degrees of freedom under the null hypothesis of no model dependence in the climate response (H_0: $\gamma_{mF} = 0$ for all models). The null hypothesis is rejected at the a% level if $F_\gamma > F_{(100-a)\%,M-1,N-2M}$, where $F_{(100-a)\%,M-1,N-2M}$ is the $(100 - a)$% quantile of the F distribution with $M - 1$ and $N - 2M$ degrees of freedom.

Similarly, F_α has a F distribution with $M - 1$ and $N - (M + 1)$ degrees of freedom under the null hypothesis of no model dependence in the historical climate (H_0: $\alpha_m = 0$ for all models). The null hypothesis is rejected at the a% level if $F_\alpha > F_{(100-a)\%,M-1,N-(M+1)}$, where $F_{(100-a)\%,M-1,N-(M+1)}$ is the $(100 - a)$% quantile of the F distribution with $M - 1$ and $N - (M + 1)$ degrees of freedom.

e. The t tests and confidence intervals

The estimates of the expected climate response $\hat{\beta}_F$ in Eqs. (A2c), (A6c), and (A10b) are linear combinations of the γ_{msr}. The y_{msr} is assumed to be normally distributed. Linear combinations of normal random variables are also normally distributed. However, σ^2 is unknown and must be estimated by s^2 in $\mathrm{Var}(\hat{\beta}_F)$. Therefore, $\hat{\beta}_F$ has a t distribution with $N - P$ degrees of freedom. Here P is the number parameters to be estimated and depends on which framework is being used for estimation.

Since $\hat{\beta}_F$ is t distributed, then T_β has a standard t distribution with $N - P$ degrees of freedom under the null hypothesis of no climate response (H_0: $\beta_F = 0$). The null hypothesis is rejected at the a% level if $T_\beta > t_{[100-(a/2)]\%,N-P}$, where $t_{[100-(a/2)]\%,N-P}$ is the $[100 - (a/2)]\%$ quantile of the t distribution with $N - P$ degrees of freedom.

A $100(1 - a)\%$ confidence interval for the actual value of the expected climate response β_F is given by

$$\begin{aligned}\hat{\beta}_F - t_{[(100-(a/2)]\%,N-P}\sqrt{\mathrm{Var}(\hat{\beta}_F)} &\le \beta_F \\ &\le \hat{\beta}_F + t_{[(100-(a/2)]\%,N-P}\sqrt{\mathrm{Var}(\hat{\beta}_F)}. \qquad \text{(A12)}\end{aligned}$$

The same theory applies to the estimates $\hat{\mu}$, $\hat{\alpha}_m$, $\hat{\gamma}_{mF}$, and $\hat{y}_{msr}$, all of which also have t distributions with $N - P$ degrees of freedom. Therefore, the significance tests on the individual model effects α_m and γ_{mF} may be conducted as above by substituting for $\hat{\beta}_F$ and $\mathrm{Var}(\hat{\beta}_F)$. The same applies to confidence intervals for the actual values of μ, α_m, γ_{mF}, and y_{msr}.

REFERENCES

Anderson, D., K. I. Hodges, and B. J. Hoskins, 2003: Sensitivity of feature-based analysis methods of storm tracks to the form of background field removal. *Mon. Wea. Rev.,* **131,** 565–573.

Bengtsson, L., K. I. Hodges, and E. Roeckner, 2006: Storm tracks and climate change. *J. Climate,* **19,** 3518–3543.

——, ——, and N. Keenlyside, 2009: Will extratropical storms intensify in a warmer climate? *J. Climate,* **22,** 2276–2301.

Bracegirdle, T. J., and D. B. Stephenson, 2012: Higher precision estimates of regional polar warming by ensemble regression of climate model projections. *Climate Dyn.,* **39,** 2805–2821, doi:10.1007/s00382-012-1330-3.

Buser, C. M., H. R. Künsch, D. Lüthi, M. Wild, and C. Schär, 2009: Bayesian multi-model projection of climate: Bias assumptions and interannual variability. *Climate Dyn.,* **33,** 849–868, doi:10.1007/s00382-009-0588-6.

Catto, J. L., L. C. Shaffrey, and K. I. Hodges, 2011: Northern Hemisphere extratropical cyclones in a warming climate in the HiGEM high-resolution climate model. *J. Climate,* **24,** 5336–5352.

Chandler, R. E., 2013: Exploiting strength, discounting weakness: Combining information from multiple climate simulators. *Philos. Trans. Roy. Soc. London,* **A371,** 1471–2962, doi:10.1098/rsta.2012.0388.

Christensen, J. H., F. Boberg, O. B. Christensen, and P. Lucas-Picher, 2008: On the need for bias correction of regional climate change projections of temperature and precipitation. *Geophys. Res. Lett.,* **35,** L20709, doi:10.1029/2008GL035694.

Collins, M., R. E. Chandler, P. M. Cox, J. M. Huthnance, J. Rougier, and D. B. Stephenson, 2012: Quantifying future climate change. *Nat. Climate Change,* **2,** 403–409, doi:10.1038/nclimate1414.

DelSole, T., 2007: A Bayesian framework for multimodel regression. *J. Climate,* **20,** 2810–2826.

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

Ferro, C. A. T., 2004: Attributing variation in a regional climate change modelling experiment. EU Project PRUDENCE Tech. Rep., 21 pp. [Available online at http://prudence.dmi.dk/public/publications/analysis_of_variance.pdf.]

Giorgi, F., and L. O. Mearns, 2002: Calculation of average, uncertainty range, and reliability of regional climate changes from AOGCM simulations via the "Reliability Ensemble Averaging" (REA) method. *J. Climate,* **15,** 1141–1158.

Hingray, B., A. Mezghani, and T. A. Buishand, 2007: Development of probability distributions for regional climate change from uncertain global mean warming and an uncertain scaling relationship. *Hydrol. Earth Syst. Sci.,* **11,** 1097–1114, doi:10.5194/hess-11-1097-2007.

Hodges, K. I., 1994: A general method for tracking analysis and its application to meteorological data. *Mon. Wea. Rev.,* **122,** 2573–2585.

——, 1995: Feature tracking on the unit sphere. *Mon. Wea. Rev.,* **123,** 3458–3465.

——, 1996: Spherical nonparametric estimators applied to the UGAMP model integration for AMIP. *Mon. Wea. Rev.,* **124,** 2914–2932.

——, 1999: Adaptive constraints for feature tracking. *Mon. Wea. Rev.,* **127,** 1362–1373.

——, R. W. Lee, and L. Bengtsson, 2011: A comparison of extratropical cyclones in recent reanalyses ERA-Interim, NASA MERRA, NCEP CFSR, and JRA-25. *J. Climate,* **24,** 4888–4906.

Hoskins, B. J., and K. I. Hodges, 2002: New perspectives on the Northern Hemisphere winter storm tracks. *J. Atmos. Sci.,* **59,** 1041–1061.

Kang, E. L., and N. Cressie, 2013: Bayesian hierarchical ANOVA of regional climate-change projections from NARCCAP Phase II. *Int. J. Appl. Earth Obs.,* **22,** 3–15, doi:10.1016/j.jag.2011.12.007.

Kharin, V. V., and F. W. Zwiers, 2002: Climate predictions with multimodel ensembles. *J. Climate,* **15,** 793–799.

Knutti, R., G. Abramowitz, M. Collins, V. Eyring, P. J. Gleckler, B. Hewitson, and L. Mearns, 2010a: Good practice guidance paper on assessing and combining multi model climate projections. Rep. of IPCC Expert Meeting on Assessing and Combining Multi Model Climate Projections, 13 pp. [Available online at http://www.ipcc-wg2.gov/meetings/EMs/IPCC_EM_MME_GoodPracticeGuidancePaper.pdf.]

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

Krzanowski, W. J., 1998: *An Introduction to Statistical Modelling.* John Wiley and Sons, 264 pp.

Livezey, R. E., and W. Y. Chen, 1983: Statistical field significance and its determination by Monte Carlo techniques. *Mon. Wea. Rev.,* **111,** 46–59.

Mastrandrea, M. D., and Coauthors, 2010: Guidance note for lead authors of the IPCC Fifth Assessment Report on consistent treatment of uncertainties. Tech. Rep., 4 pp. [Available online at https://www.ipcc-wg1.unibe.ch/guidancepaper/ar5_uncertainty-guidance-note.pdf.]

McDonald, R. E., 2011: Understanding the impact of climate change on Northern Hemisphere extra-tropical cyclones. *Climate Dyn.,* **37,** 1399–1425, doi:10.1007/s00382-010-0916-x.

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

——, C. Covey, K. E. Taylor, T. Delworth, R. J. Stouffer, M. Latif, B. McAvaney, and J. F. B. Mitchell, 2007b: The WCRP CMIP3 multimodel dataset: A new era in climate change research. *Bull. Amer. Meteor. Soc.,* **88,** 1383–1394.

Moss, R. H., and Coauthors, 2010: The next generation of scenarios for climate change research and assessment. *Nature,* **463,** 747–756, doi:10.1038/nature08823.

Peña, M., and H. van den Dool, 2008: Consolidation of multimodel forecasts by ridge regression: Application to Pacific sea surface temperature. *J. Climate,* **21,** 6521–6538.

Pennell, C., and T. Reichler, 2011: On the effective number of climate models. *J. Climate,* **24,** 2358–2367.

Räisänen, J., 2001: CO_2-induced climate change in CMIP2 experiments: Quantification of agreement and role of internal variability. *J. Climate,* **14,** 2088–2104.

Sain, S. R., D. Nychka, and L. Mearns, 2011: Functional ANOVA and regional climate experiments: A statistical analysis of dynamic downscaling. *Environmetrics,* **22,** 700–711, doi:10.1002/env.1068.

Sanderson, B. M., and R. Knutti, 2012: On the interpretation of constrained climate model ensembles. *Geophys. Res. Lett.,* **39,** L16708, doi:10.1029/2012GL052665.

Solomon, S., D. Qin, M. Manning, Z. Chen, M. Marquis, K. Averyt, M. Tignor, and H. L. Miller Jr., Eds., 2007: *Climate Change 2007: The Physical Science Basis.* Cambridge University Press, 996 pp.

Stephens, M. A., 1974: EDF statistics for goodness of fit and some comparisons. *J. Amer. Stat. Assoc.,* **69,** 730–737, doi:10.2307/2286009.

Stephenson, D. B., M. Collins, J. C. Rougier, and R. E. Chandler, 2012: Statistical problems in the probabilistic prediction of climate change. *Environmetrics,* **23,** 364–372, doi:10.1002/env.2153.

Taylor, K. E., R. J. Stouffer, and G. A. Meehl, 2012: An overview of CMIP5 and the experiment design. *Bull. Amer. Meteor. Soc.,* **93,** 485–498.

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

——, J. M. Arblaster, and R. Knutti, 2011a: Mapping model agreement on future climate projections. *Geophys. Res. Lett.,* **38,** L23701, doi:10.1029/2011GL049863.

——, B. Sansó, and R. L. Smith, 2011b: Characterizing uncertainty of future climate change projections using hierarchical Bayesian models. *Bayesian Statistics 9,* J. M. Bernardo et al., Eds., Oxford University Press, 706 pp.

Ventura, V., C. J. Paciorek, and J. S. Risbey, 2004: Controlling the proportion of falsely rejected hypotheses when conducting multiple tests with climatological data. *J. Climate,* **17,** 4343–4356.

Weigel, A. P., R. Knutti, M. A. Liniger, and C. Appenzeller, 2010: Risks of model weighting in multimodel climate projections. *J. Climate,* **23,** 4175–4191.

Woollings, T., J. M. Gregory, J. G. Pinto, M. Reyers, and D. J. Brayshaw, 2012: Response of the north atlantic storm track to climate change shaped by oceanatmosphere coupling. *Nat. Geosci.,* **5,** 313–317, doi:10.1038/ngeo1438.

Yip, S., C. A. T. Ferro, D. B. Stephenson, and E. Hawkins, 2011: A simple, coherent framework for partitioning uncertainty in climate predictions. *J. Climate,* **24,** 4634–4643.

Zappa, G., L. C. Shaffrey, and K. I. Hodges, 2013a: The ability of CMIP5 models to simulate North Atlantic extratropical cyclones. *J. Climate,* in press.

——, ——, ——, P. G. Sansom, and D. B. Stephenson, 2013b: A multimodel assessment of future projections of North Atlantic and European extratropical cyclones in the CMIP5 climate models. *J. Climate,* in press.

Zwiers, F. W., 1987: A potential predictability study conducted with an atmospheric general circulation model. *Mon. Wea. Rev.,* **115,** 2957–2974.

——, 1996: Interannual variability and predictability in an ensemble of AMIP climate simulations conducted with the CCC GCM2. *Climate Dyn.,* **12,** 825–847, doi:10.1007/s003820050146.