Journal of Climate

A simple framework for weighting climate change projections in multi-model ensembles

--Manuscript Draft--

Manuscript Number:	
Full Title:	A simple framework for weighting climate change projections in multi-model ensembles
Article Type:	Article
Corresponding Author:	Philip George Sansom, BSc University of Exeter Exeter, Devon UNITED KINGDOM
Corresponding Author's Institution:	University of Exeter
First Author:	Philip George Sansom, BSc
Order of Authors:	Philip George Sansom, BSc
	David B. Stephenson
	Christopher A. T. Ferro
	Giuseppe Zappa
	Len C. Shaffrey
Abstract:	Future climate projections are often based on analysis of multi-model ensembles of simulations from global climate models. Current methods for synthesising data from multi-model ensembles are largely heuristic in nature. This paper introduces a nested family of three simple statistical frameworks for analysing climate projections from multi-model ensembles. This family includes frameworks that yield the ``one model, one vote'' and ``one run, one vote'' weighting approaches often used in climate projection. In addition, it is shown that an intermediate framework exists which is optimal when there is good agreement between models on the climate response. The ``one model, one vote'' approach is not optimal when there is good agreement on the climate response. The ``one run, one vote'' approach is only valid when there is also good agreement between models on the historical climate. Significance tests are derived to choose the most appropriate framework for specific multi-model ensemble data. The assumptions underlying each framework are explicit and can be checked using simple tests and graphical techniques. The frameworks can be used to test for significant evidence of non-zero climate response and to construct confidence intervals for the size of the response. The methodology is illustrated using data for the North Atlantic storm track from the CMIP5 multi-model ensembles. Good agreement is found on the future change in frequency of cyclones over most of the region, but not on the historical frequency. Significant decreases in cyclone frequency are found on the flanks of the main North Atlantic storm track and in the Mediterranean basin.
Suggested Reviewers:	

Page and Color Charge Estimate Form

Click here to download Page and Color Charge Estimate Form: Page-Charge-Estimates.pdf

A simple framework for weighting climate change projections in multi-model ensembles

Philip G. Sansom *

University of Exeter, Exeter, United Kingdom

David B. Stephenson

University of Exeter, Exeter, United Kingdom

Christopher A. T. Ferro

University of Exeter, Exeter, United Kingdom

Giuseppe Zappa

University of Reading, Reading, United Kingdom

Len Shaffrey

University of Reading, Reading, United Kingdom

* *Corresponding author address:* Philip Sansom, Harrison Building, University of Exeter, North Park Road, Exeter, EX4 4QF, United Kingdom.

E-mail: pgs201@exeter.ac.uk

ABSTRACT

Future climate projections are often based on analysis of multi-model ensembles of simulations from global climate models. Current methods for synthesising data from multi-model ensembles are largely heuristic in nature. This paper introduces a nested family of three simple statistical frameworks for analysing climate projections from multi-model ensembles.

This family includes frameworks that yield the "one model, one vote" and "one run, one vote" weighting approaches often used in climate projection. In addition, it is shown that an intermediate framework exists which is optimal when there is good agreement between models on the climate response. The "one model, one vote" approach is not optimal when there is good agreement on the climate response. The "one run, one vote" approach is only valid when there is also good agreement between models on the historical climate.

Significance tests are derived to choose the most appropriate framework for specific multi-model ensemble data. The assumptions underlying each framework are explicit and can be checked using simple tests and graphical techniques. The frameworks can be used to test for significant evidence of non-zero climate response and to construct confidence intervals for the size of the response.

The methodology is illustrated using data for the North Atlantic storm track from the CMIP5 multi-model ensembles. Good agreement is found on the future change in frequency of cyclones over most of the region, but not on the historical frequency. Significant decreases in cyclone frequency are found on the flanks of the main North Atlantic storm track and in the Mediterranean basin.

1. Introduction

Future climate projections are usually derived using simulations from Global Climate Models (GCMs). The previous phase of the World Climate Research Programme (WCRP) Coupled Model Intercomparison Project (CMIP3) included 24 models from 17 groups in 12 countries (Meehl et al. 2007). The latest CMIP5 multi-model ensemble (MME) (Taylor et al. 2012) is not yet fully populated but promises to include an even greater number of more recent models. 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 synthesise it are urgently needed".

MMEs are useful for exploring different sources of uncertainty in climate projections. An MME will usually include projections of multiple future emissions scenarios. Ideally several runs of each scenario should be available from each model. The different models, scenarios and runs explore the three primary sources of uncertainty in climate projections. The different runs sample the internal variability of the climate by perturbing the initial conditions. The different scenarios represent our uncertainty about changes in radiative forcing due to future emissions. These depend on complex socio-economic factors that are difficult to predict. Any projection is therefore conditional on the emissions scenario being modelled. Model uncertainty arises from the fact that not all relevant processes are well represented by climate models.

The projections presented in the IPCC Fourth Assessment Report (Solomon et al. 2007) were largely based on equally weighted multi-model means of the projections from the models in the CMIP3 MME. The equally weighted multi-model mean treats all models as equally credible i.e. "one model, one vote" (Knutti et al. 2010a). When multiple runs are available from a model these are first averaged together before averaging over all models. This approach reduces the ensemble of projections to a single point estimate.

The equally weighted multi-model mean is a heuristic point estimate that has a number of shortcomings. The assumptions underlying the estimate are not explicit and therefore

cannot be checked. If the assumptions cannot be checked then there is no guarantee that the estimate is meaningful. The equally weighted multi-model mean is also not resistant to outliers. Runs which are outlying compared to the rest of the MME may strongly influence the estimate. Furthermore, no attempt is made to quantify the uncertainty in the estimate. All of these shortcomings may be overcome by specifying our assumptions about the structure of the uncertainty in the MME using a statistical framework. Once the assumptions are specified explicitly they can be checked. If the structure of the uncertainty is specified then it can be estimated and confidence intervals can be constructed. Also, once the structure of the uncertainty is specified, outlying runs can be systematically identified.

One alternative to the "one model, one vote" approach is to treat each run equally rather than each model i.e. "one run, one vote". If the models all perform similarly this estimate may appear sensible. If not, there is a risk that a poorly performing model is given too much weight due to having more runs available. Although both approaches have been utilised in climate science, no clear statement has been made of the assumptions underlying each approach or when one approach might be more appropriate than the other.

Alternative approaches have been proposed which assign greater weight to models perceived as performing well according to some metric (see Knutti et al. (2010b); Collins et al. (2012) and references therein). Many of these approaches are heuristic in nature while some derive their weights from hierarchical Bayesian frameworks. As yet, there exists little agreement on the best performance metrics to use in these frameworks or even structure of the frameworks themselves (Stephenson et al. 2012). Hierarchical Bayesian frameworks are also extremely complex and require specialist knowledge to implement and interpret.

In this study, analysis of variance (ANOVA) frameworks are used to make explicit the assumptions underlying the "one model, one vote" and "one run, one vote" approaches to estimating the expected climate response in a MME. ANOVA frameworks have been used in climate science for a variety of purposes (Zwiers 1987, 1996; Räisänen 2001). Simple ANOVA frameworks have already been applied to analysis of MMEs of 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 2012).

Recently, Yip et al. (2011) used a simple two-way ANOVA framework to partition the uncertainty in the CMIP3 MME. Structural differences result in models simulating different climate responses to the same emissions scenarios. Using the ANOVA framework Yip et al. (2011) were able to quantify this "interaction" between models and emission scenarios. This study shows how ANOVA frameworks can also be used to estimate the size of the individual effects relating to the various factors included in the framework e.g. the difference between the climates simulated in historical and future scenarios.

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 applied to the CMIP5 MME to analyse the future climate response of the North Atlantic storm track.

2. Statistical Frameworks

This section begins with the general form of a multi-model mean estimate of the climate response in an MME. A family of ANOVA frameworks are then outlined, including cases equivalent to the "one model, one vote" and "one run, one vote" multi-model means. The rest of the section is devoted to statistical inference within these frameworks. This includes verifying the underlying assumptions, choosing the most appropriate framework and the construction of significance tests and confidence intervals.

105 *a. The multi-model mean response*

Let y_{msr} represent a climate statistic (e.g. a 30 year 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 mean climate in the future and historical scenarios

$$\overline{y}_{mF.} - \overline{y}_{mH.} \tag{1}$$

where $\overline{y}_{ms.}$ is the 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 multi-model mean estimate of the climate response is given by

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

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 multi-model mean, i.e. the "one model, one vote" approach, where

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

alternatively the "one run, one vote" approach has weights

$$W_{mH} = R_{mH} \quad \text{and} \quad W_{mF} = R_{mF} \tag{4}$$

106 *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 Eqn. 3 is equivalent to the maximum-likelihood (ML) estimate $\hat{\beta}_F$ of the ex-

pected climate response β_F from the following two-way ANOVA framework with interactions

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

with the usual constraints that $\sum_{m=1}^{M} \alpha_m = 0$, $\beta_H = 0$, $\gamma_{mH} = 0$ for all models and $\sum_{m=1}^{M} \gamma_{mF} = 0$. The effect μ is the expected climate in the historical scenario and β_F is the expected climate response between scenarios F and H. The effect α_m is the difference of climate of model m from the expected climate under the historical scenario. The interaction terms γ_{mF} represent the difference between the climate response simulated by model m and the expected climate response β_F.

The μ, α_m, β_s and γ_{ms} effects are all assumed to be fixed. We do not consider how the effects might vary if we were to use a different set of climate models. The random component ϵ_{msr} represents the internal variability of y_{msr} and is assumed to be normally distributed. The central limit theorem implies that any long term mean will be approximately normally distributed (if the climate response trend is small).

There are a total of $2M$ parameters to be estimated in the two-way ANOVA framework of Eqn. 5. One parameter must be estimated for the expected historical climate μ and one for the expected climate response β_F. The α_m and γ_{mF} effects are constrained to be centered on μ and β_F respectively. Therefore only $M - 1$ of each need 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. In that case all the degrees of freedom are 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.

c. An simpler additive ANOVA framework

The interaction effects in Eqn. 5 allow for the possibility that each model responds differently to the future scenario. If the models all respond similarly then these parameters are unnecessary. Estimating a systematic component where none exists introduces bias, which leads to decreased precision in the estimates. A more parsimonious additive framework may be more appropriate

$$\begin{aligned} y_{msr} &= \mu + \alpha_m + \beta_s + \epsilon_{msr} \\ \epsilon_{msr} &\overset{iid}{\sim} N\left(0, \sigma^2\right) \end{aligned} \tag{6}$$

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 Eqn .5. However the maximum likelihood estimates of the effects are not the same.

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

$$W_{mH} = W_{mF} = \frac{R_{mH} R_{mF}}{R_{mH} + R_{mF}} \tag{7}$$

The weights depend on the number of runs available from each model but are not equivalent to the "one run, one vote" estimate. Knutti et al. (2010a) advise against "inappropriately" weighting models based on the number of runs they contribute to the MME. This additive framework assumes that all models simulate the same climate response with the same internal variability. Therefore if that assumption is believable then we should give increased weight to models that have more runs. Note however 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.

This additive framework has only $M + 1$ parameters to be estimated. Without the inter-

action effects there are $M-1$ less parameters to be estimated. An additional $M-1$ degrees of freedom are then available to estimate the uncertainty. Therefore the precision of the parameter estimates should increase compared to the two-way framework with interactions. However if the models do not all respond similarly then a systematic component is missing from the framework. The precision of the estimates will decrease dramatically if the missing effects are large. The simple additive framework must therefore only be used when there is good agreement between models on the climate response.

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 similar historical climates then a one-way ANOVA framework is be more appropriate

$$\begin{aligned} y_{msr} &= \mu + \beta_s + \epsilon_{msr} \\ \epsilon_{msr} &\overset{iid}{\sim} N\left(0, \sigma^2\right) \end{aligned} \tag{8}$$

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

In the Appendix it is shown that the maximum likelihood estimate $\hat{\beta}_F$ of the expected climate response from this one-way framework is equivalent to the "one run, one vote" estimate of Eqn. 4. If the assumptions of the on-way framework are believable then each run is weighted equally.

This simple framework has only two parameters to be estimated. With more 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 simulate similar historical climates and climate responses the precision of the estimates may decrease dramatically. This simple framework must therefore only by used if both these assumptions are satisfied.

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 i.e. the model-scenario means $\overline{y}_{ms.}$. This simplicity comes at the cost of requiring a balanced design i.e. 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 which do not have sufficient runs. This can be avoided by fitting the ANOVA framework using normal linear regression methods.

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

These assumptions must be carefully checked before confidence can be placed in the estimates from the frameworks. If these they are satisfied then the ANOVA framework provides a compact statistical description of the MME. If required, the ANOVA framework could be used to generate a new ensemble of runs which should be statistically indistinguishable from repeating the original ensemble of runs for the same scenarios with the same models i.e. the ANOVA framework is an emulator for the entire MME.

The assumption of independence is difficult to verify so consideration must be given *a priori* to whether this assumption is justified. The distributional assumptions may be verified 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 standardised 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 standardised residuals against the fitted

values $\hat{y}_{msr}$ should show random scatter about zero. Any systematic component visible in the scatter may indicate non-constant variance or a systematic difference between the y_{msr} that is not captured by the framework.

f. Identifying outlying runs

The ϵ_{msr} are assumed to be normally distributed. Therefore fitted residuals e_msr should also be normally distributed. Any runs having standardised fitted residuals lying in the far tails of the standard normal distribution are considered outlying. If viewed as a significance test we might consider labelling any run with a standardised 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% criteria ($|Z| > 2.58$) is therefore more appropriate.

Outliers can be easily identified from the plot of standardised residuals against observed values y_{msr}. They may also be visible in the quantile-quantile plot used in the check for normality. This procedure provides an objective approach to identifying potentially problematic climate model runs. However, 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 ensemble unless an explanation can be found for the unusual behaviour.

Outlying runs may 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 this case the outlying runs do not effect the analysis and can remain in the ensemble. If removing the outliers strongly effects the estimates of the main effects μ and β_F it is essential to determine whether the outlying runs represent plausible climates or poor simulations.

Outlying runs may also affect the check 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 which unlikely given the total number of runs. This can quickly be checked by *temporarily* remove 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, possibly after removing further outliers, an ANOVA framework may not be appropriate.

g. Which framework is most appropriate?

In Section c it is noted that the additive framework is only appropriate if all models simulate the same climate response. Similarly in Section d it is noted that the additive 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. If the expected effect size 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 significance test is derived to test for the presence of model-dependent responses i.e. to test 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 and additive frameworks 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 model-dependent climate responses to that explained by internal variability. If the p-value of the test is small ($p < a$) we conclude that there is significant evidence of model-dependent climate response at the $a\%$ level and

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 of where $\alpha_m = 0$ for all models m. In the Appendix, a significance test is derived to test for the presence of model specific discrepancies i.e. to test null hypothesis $H_0 : \alpha_m = 0$ for all m against the alternative $H_a : \alpha_m \neq 0$ for any m. The test statistic is the ratio of variances

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

R_β^2 is the coefficient of determination under the one-way model of Eqn. 8. The quantity f_α^2 represents the ratio of the variance explained by model-specific discrepancies to that explained by internal variability. If the p-value of the test is small ($p < a$) we conclude that there is significant evidence of model-specific discrepancy at the $a\%$ level and 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 is 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 i.e. to test the null hypothesis $\beta_F = 0$ against the alternative $\beta_F \neq 0$. The test statistic is:

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

If the p-value of the test is small ($p < a$) then we conclude that there is significant evidence of a non-zero climate response at the $a\%$ level of significance. If the p-value is not small then 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 i.e. $d_\beta \simeq 2$ implies the

projected future climate is more extreme than 95% of plausible historical climates. The IPCC Fourth Assessment Report (Meehl and Coauthors 2007) highlights climate responses greater than one standard deviation in magnitude i.e. $d_\beta > 1$. 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 focussed studies.

i. Testing of individual climate models

Similar tests to that given for non-zero expected climate response in the previous section can be made for non-zero model-dependant climate response ($\gamma_{mF} \neq 0$) and non-zero model-specific discrepancy ($\alpha_m \neq 0$) in individual models. Under the null hypotheses of no model-specific discrepancy ($H_0 : \alpha_m = 0$) and no model-dependant climate response ($H_0 : \gamma_{mF} = 0$) the test statistics are:

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

If the p-value of one of these tests is small ($p < a$) we conclude that there is significant evidence at the a% level of model-specific discrepancy or model specific climate response respectively. This means that particular model does not agree with the expected historical climate or expected climate response in the MME. Removing models which do not agree with the expected climate or climate response may lead to a more homogeneous MME. This may come at the expense of not sampling unlikely yet still plausible climates.

j. Is the ensemble large enough?

The weak law of large numbers states that the sample mean converges towards the expected value as the sample size increases. Thus for models or MMEs with a large number of runs we can be confident that the estimates from the ANOVA frameworks approach their expected values. However, large numbers of runs are rarely available.

The power of a hypothesis test is the probability that the null hypothesis is rejected given

that it is not true i.e. the probability that we conclude $\beta_F \neq 0$ given that the expected value $\beta_F = \beta \neq 0$. The power depends on the size of the test (the level of significance a at which the null hypothesis is rejected), the amount of data (i.e. the number of runs), the size of the expected effect and the size of the internal variability. However the internal variability is unknown and must also be estimated.

For all of frameworks we have discussed the parameter estimates are weighted linear combinations of the y_{msr}. Therefore the variances of the parameter estimates are proportional to the size σ^2 of the internal variability ϵ_{msr}. By specifying the standardized effect size d_β we avoid having to specify the internal variability directly. Figure 1(a) illustrates the power of a test of size 10% to detect climate responses of various sizes d_β in MMEs of different sizes.

A similar analysis can be performed for the tests for model-dependent climate response and model-specific discrepancy. In this case the expected effect size is specified in terms of f_γ^2 or f_α^2. Figures 1(b) and 1(c) illustrate the power of tests of size 10% to detect overall model-dependent climate response or model-specific discrepancy effects of various sizes in MMEs of different sizes.

k. 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 Eqn. 2 using the weights in Eqns. 3, 7 and 4. 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 an 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:
   (a) If the assumptions appear satisfied and there are no outlying runs then go to the next step.
   (b) If there are outlying runs, investigate possible causes before removing completely, or consider removing *temporarily* and re-checking the assumption of normality.
   (c) 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-dependent climate response. If the null hypothesis of no model-dependent response is rejected then stop, the two-way framework is the 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-specific discrepancy. If the null hypothesis of no model-specific discrepancy is rejected then stop, the additive framework is the 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 non-zero 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, 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 track density of extra-tropical cyclones from an ensemble of climate models participating in the WCRP CMIP Phase 5 (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. (2012a). Six hourly output suitable for storm track analysis is available from 19 models from 12 centers. Projections are compared from two 30 year periods. The recent climate is represented by the mean of a 30 year period from the historical experiment between December 1975 and January 2005. The future climate is analysed conditional on the RCP4.5 midrange mitigation emissions scenario (Moss et al. 2010). The mean of a 30 year period between December 2070 and January 2100 is analysed. At least one realisation is available from each model for each scenario. The total number of realisations available for each model- scenario pair is

summarised in Table 2.

The analysis methodology is similar to that used in several previous studies of extra-tropical cyclones (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 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 two days or travel less than 1000km are excluded. Spatial statistics are then computed from the tracks using the spherical kernel approach of Hodges (1996).

This study focuses on the track density statistics. This is the mean number of cyclones passing a particular point each month. The spherical kernel approach utilises a variable bandwidth so the statistics are rescaled to be representative of a region of 5 degrees in width centred on a particular grid point. This study focuses on the December-January-February (DJF) winter period in the North Atlantic. The study region is defined as 80E to 40W and 30N to 90N. 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. 2. The CMIP5 models appear to simulate the DJF storm track reasonably well but with some departures. The main north-east track is too weak while the more zonal track towards northern Europe is too strong. Comparing the climate response estimates from the three frameworks in Figs. 2(a), (b) and (c) suggests the additive framework may be suitable to describe the CMIP5 MME. The response estimates from the two-way and additive frameworks appear

similar. The response estimate from the one-way framework fails to capture the increase in track density over the UK and Denmark indicated by the other two frameworks. This suggests the presence of large discrepancies between the historical climates simulated by the CMIP5 models.

2) A single grid point

To better understand the differences between the climate response estimates, a single grid point in central France (46.5°N 1.25°E) is considered in detail. Fig. 3 confirms that there are large differences between the historical climates simulated by the CMIP5 models. By comparison, the usual unweighted multi-model mean 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. An exception is the MIROC-ESM model which appears to take 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 due to internal variability.

The differences between the structures of the three ANOVA frameworks are also visible in Fig. 3. In the two-way framework the maximum-likelihood estimate of the mean climate in each model and each scenario is the unweighted mean of the runs from that model and that scenario (Fig. 3(a)). 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. 3(b)). As expected, the uncertainty indicated by the error bars is reduced compared to the two-way framework. In most cases the error bars include the majority of

runs from each model and scenario. This gives confidence that the estimates are reasonable.

The one-way framework constrains the estimates so that all models simulate the same same historical climate and climate response (Fig. 3(c)). Note that these do not coincide with the usual unweighted multi-model mean estimates. The error bars indicate that the uncertainty is greater in the one-way framework than the additive framework. This is not surprising given the large differences between the historical climate 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.

No systematic patterns are visible in the plot of standardised residuals against the fitted values from the two-way framework in Fig. 4(a). 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. 4(b). The majority of the runs lie close to the expected straight line, although some skewness is visible. This is likely to be due to the influence of the two outliers. After removing the two runs of MIROC-ESM no further outliers are identified and the skew in the quantile-quantile plot is reduced (not shown). The p-value of the Anderson-Darling test for normality is 0.19 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 reassured that the framework assumptions are basically justified at this grid point.

The variance ratio f_γ^2 is calculated as 0.33, i.e. differences in the climate response between models are responsible for variability equivalent to 33% of that explained by internal variability. The p-value of the significance test for model-dependent climate response based on f_γ^2 is 0.758. 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 46.6, i.e. differences between the historical climates simulated by the models explain 47 times more variation in the CMIP5 MME than the internal variability. This result is highly significant. The null hypothesis of no model-specific discrepancies is rejected entirely. At this grid point, the additive framework provides the most parsimonious description of the MME.

3) The North Atlantic storm track

Figure 2 suggests that while the CMIP5 models simulate different historical climates, they agree reasonably well on the future climate response. Before the hypothesis of no model-dependent response can be tested the framework assumptions must be checked under the two-way framework.

Examining plots of standardised residuals against fitted values at a random selection of grid points reveals no evidence of non-constant variance between models or scenarios. One of the two runs of the MIROC-ESM model identified as outlying over central France is also identified as an outlier at many other grid points. The Anderson-Darling test (Fig. 5) also suggests that the assumption of normality is not well justified over parts of western and northern Europe. To investigate these artefacts further the standardised residuals of each run were mapped individually ($N = 78$ plots, not shown). The second run of the historical scenario from MIROC-ESM is identified as outlying at the 1% level at 25.6% of grid points, spread widely over the study region. No other run is identified at more than 6.4% of grid points.

After removing the second historical run of the MIROC-ESM no single run is identified as outlying at the 1% level at more than 4.4% of grid points. The Anderson-Darling test (not shown) indicates only a few scattered points where the assumption of normality may not be justified. Investigating the reasons for apparently outlying run of MIROC-ESM is beyond scope of this study. 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 reassured that the framework assumptions are basically justified across the study region.

The variance ratio f_γ^2 and p-values of the significance test for model-dependent climate response are shown in Figs. 6(a) and (b). The uncertainty due to model-dependent climate response is less than that due to internal variability over most of the study region. However, areas of significant non-zero model dependence at the 10% level are detected, most notably over the sub-tropical Atlantic ocean.

In order 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 (Eqn. 12) are mapped in Fig. 7. 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 sub-tropical Atlantic ocean CSIRO-Mk3.6.0, FGOALS-g2, MIROC-ESM and MIROC-ESM-CHEM all deviate strongly from the ensemble mean response. MRI-CGCM3 is unique in that it deviates from the ensemble mean near the Iberian peninsula but not over the rest of the sub-tropical Atlantic.

Figure 7 indicates that all the regions of interaction detected in Fig. 6(b) involve more than one model. Comparing plots in Fig. 7 shows that models which 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 climates in other regions.

Although there is evidence of model-dependence in the climate response in some regions, there appears to be reasonable agreement in others. Where there is agreement on the response, 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 discrepancies in their historical climates. Differences between the historical climates of the models explain at least twice the amount of variation explained by the internal variability, everywhere in the study region. Over central Europe the variance ratio rises to $f_\alpha^2 \approx 60$. This agrees with Zappa et al. (2012b) who found that the storm tracks of several models extend too far into the European continent. Based on 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.

For simplicity one might wish to use only one framework over the whole study region. However, comparing Fig. 8(b) with Fig. 6(b) shows that the mean climate response estimate from the additive framework has greater precision where there is no significant evidence of model-dependence. Note that the decrease in precision from using the two-way framework where there is no model-dependence is generally small compared to the decrease in precision from using the additive framework where there is model-dependence. Therefore if only one framework is used the two-way framework should be preferred. This agrees with the theoretical arguments in Section c.

The difference between the estimates of the expected climate response β_F from the two-way and additive frameworks is shown in Fig. 8(a). A comparison with Figs. 2(a) and (b) shows that the two-way framework 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 Tab. 2 the additive framework assigns most weight to the CSIRO-Mk3.6.0, EC-EARTH, IPSL-CM5A-LR and MPI-ESM-LR models. Comparing these models in Fig. 3(a) shows that they all have relatively weak climate responses. Since the additive framework up-weights these models, its climate response estimate is correspondingly lower.

Figures 6(c) and (d) compare the standardised climate responses d_β from the two-way and additive frameworks. In many regions the standardized response estimated by the two-way framework is greater than that estimated by the additive framework. This includes some regions where there is no significant evidence of model-dependence in the response. In

this case the decrease in the estimate of the expected climate response β_F from the two-way framework to the additive framework is greater than the increase in the precision of the estimate of the internal variability s.

Both the two-way and additive frameworks estimate large ($d_\beta > 1$) climate responses in the sub-tropical Atlantic, the Mediterranean and parts of the main north-east branch of the storm track. The statistical significance of these responses is shown in Figs. 6(e) and (f). Both frameworks find significant evidence of non-zero climate response at the 1% level over the three regions already highlighted but also 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 both the northern and southern flanks of the North Atlantic storm track in the RCP4.5 scenario. A small increase in frequency is indicated in the zonal branch of the storm track directed towards Northern Europe. There is evidence at the 10% level but not at the 1% level that this increase is due to a change in radiative forcing rather than simply 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. Discussion and conclusions

This study has described a family of simple ANOVA frameworks which naturally yield the "one model, one vote" and "one run, one vote" estimates of future climate response in a MME. The assumptions of these frameworks may be checked using simple tests and graphical techniques. In addition to the usual point estimates confidence intervals may be constructed. ANOVA frameworks usually require a balanced number of runs to be available

for each model-scenario pair (Krzanowski 1998). This restriction is relaxed here by fitting the estimates using normal linear regression methods rather than traditional ANOVA techniques.

In the example of the North Atlantic storm track it was demonstrated how runs and models may be identified which differ from the expected behaviour in the MME. Such outliers should not simply be removed from the MME without further investigation. Outliers may arise for a variety of reasons and can be informative. These include computational or set-up errors in an individual run, or inadequate representation of certain physical processes in a particular model. Outliers may also represent unlikely but still plausible climates which contribute valuable information to the MME. The challenge is to determine which of these two cases applies. This may involve detailed process-based evaluation, comparison with historical observations, or other checking procedures.

In addition to the "one model, one vote" and "one run, one vote" approaches this study has shown that there is an intermediate case, made explicit by the additive ANOVA framework. The additive framework provides more precise estimates of the expected climate response than the two-way framework with interactions when the models all simulate similar climate responses. The model weights in this approach depend on the number of runs from each model under both the historical and future scenarios. Having many runs from only one scenario is not sufficient to achieve a high weighting. This emphasises the need for modelling centers to provide multiple runs from each future scenario, not just the historical scenario.

The two-way ANOVA with interactions shows that the equally weighted multi-model mean of the "one model, one vote" approach implicitly allows for the possibility that each climate model may respond differently to the same radiative forcing. Even if model-dependent effects are not present it is an unbiased estimate of the expected climate response in a MME. The decrease in precision associated with estimating the additional model-dependent effects should also be minimal since the estimated model-dependent effects should be small. The equally weighted multi-model mean may therefore be regarded as a conservative estimate of

the expected climate response in a MME. If additional analysis is not possible, the use of the equally weighted multi-model mean and associated confidence intervals from the two-way ANOVA with interactions is advisable.

However, the presence of the model-dependent response effects complicate the interpretation of the expected climate response. Apparent model-dependent responses may arise due to actual differences between climate models or due to internal variability in the climate system. These two situations may be distinguished using the F-tests provided by the ANOVA frameworks. But while climate models may share a common goal, they are only numerical approximations and are unlikely to ever agree completely. If the model-dependent components of the climate response γ_{mF} are small compared to the expected response β_F then the differences may not be considered important. However if the model-dependent effects are of a similar size to the expected response then it is difficult to interpret the expected response as representative of all models in the MME. The variance ratio f_γ^2 provides a dimensionless scale to quantify the level of agreement on the climate response.

When the agreement on the climate response is poor, the challenge is to determine the scientific reasons for the differences between the models. Ideally these should be understood in terms of physical processes. In some cases the climate response simulated by a particular model may depend strongly on the historical climate in that model. Such emergent constraints may also be used to weight models by estimating the relationship and comparing it to the actual climate (Bracegirdle and Stephenson (2012) and references therein).

The ANOVA frameworks in this study only describe the uncertainty in the MME and do not consider the actual climate explicitly. It is assumed that the models in the MME are centered on some expected climate. If it is believed that the models are centered on the actual climate, then the projections from the MME may be interpreted as representative of the actual climate. However, models are only approximations and there is no guarantee that the MME is centered on the actual climate. In that case, an additional source of uncertainty exists due to the discrepancy between the expected climate of the MME and the actual

climate. However, this component of uncertainty is not included explicitly in the ANOVA frameworks. Therefore, the frameworks described here will underestimate the uncertainty in the actual climate response in the presence of such a discrepancy.

Until recently it has been common to assume that the models in an MME are centered on the actual climate i.e. "truth centered". However, there is an increasing body of evidence that common biases may exist amongst GCMs Knutti et al. (2010a). The existence of shared errors is quite plausible 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).

An alternative to the "truth centered" approach is an assumption of "exchangeability" between models and the actual climate. Under this assumption, models are assumed to be equally credible and drawn from the same space of plausible climates as the actual climate (Knutti et al. 2010a). However, it may be optimistic to assume direct exchangeability between climate models and the actual climate. Rougier et al. (2012) extends the exchangeable approach to explicitly include the shared error and a scaling factor between models and the actual climate. Under this framework the climate models are assumed to be independently distributed about some ensemble mean which differs from the actual climate by the shared error.

A number of authors have proposed unifying frameworks (Tebaldi et al. 2005; Chandler 2011; Rougier et al. 2012) which include the actual climate explicitly. Some include a shared error, others do not. Tebaldi et al. (2005) assign weights to the models based on bias from the actual historical climate and convergence to the ensemble mean response (Giorgi and Mearns 2002). However, in the presence of shared errors, neither of these criteria is necessarily an indicator of good performance. Chandler (2011) and Rougier et al. (2012) therefore treat all models equally. All these approaches rely on complex Bayesian hierarchical frameworks.

Outliers and model weighting in the presence of shared discrepancies highlight the need for process-based comparisons of climate models. At present there is no consensus on a

"correct" framework for analysing MME. The ANOVA frameworks presented here are easily understood, implemented and interpreted but fail to capture any uncertainty due to shared errors between models. The complex Bayesian hierarchical frameworks developed recently aim to capture this additional uncertainty but require specialist knowledge and skills to implement. However, this should not be a barrier to adoption if this the existence of this additional source of uncertainty is accepted. Both approaches also fail to fully incorporate the expert knowledge of climate scientists to determine which models provide the best information. Only through increased co-operation between statisticians and climate scientists can these issues be resolved.

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 modelling 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

a. Derivation of two-way framework with interactions

The log-likelihood of the two-way framework with interactions in Equation 5 is

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

with the usual constraints $\sum_{m=1}^{M}\alpha_m = 0$, $\beta_H = \gamma_{mH} = 0\ \forall\ m$ and $\sum_{m=1}^{M}\gamma_{mF} = 0$. Maximum likelihood estimates are obtained by minimising 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 0. Solving the set of simultaneous equations yields the following estimates:

$$\hat{\mu} = \frac{1}{M}\sum_{m=1}^{M}\overline{y}_{mH.} \quad \text{(A2a)}$$

$$\hat{\alpha}_m = \overline{y}_{mH.} - \hat{\mu} \quad \text{(A2b)}$$

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

$$\hat{\gamma}_{mF} = \left(\overline{y}_{mF.} - \overline{y}_{mH.}\right) - \hat{\beta}_F \quad \text{(A2d)}$$

and

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

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:

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

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

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

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

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

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

b. Derivation of additive framework

The log-likelihood of the additive framework in Equation 6 is

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

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:

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

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

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

where:

$$W_m = \frac{R_{mH} R_{mF}}{R_{mH} + R_{mF}} \tag{A7}$$

The variances of the estimates are given by:

$$Var\left(\hat{\mu}\right) = \frac{\sigma^2}{M^2}\left(\sum_{m=1}^{M}\frac{1}{R_{m.}} + \frac{1}{\sum_{m=1}^{M} W_m}\left(\sum_{m=1}^{M}\frac{R_{mF}}{R_{m.}}\right)^2\right) \tag{A8a}$$

$$\begin{aligned} Var\left(\hat{\alpha}_m\right) = {} & \frac{\sigma^2}{\sum_{m=1}^{M} W_m}\left(\frac{R_{mF}}{R_{m.}} - \frac{1}{M}\sum_{m=1}^{M}\frac{R_{mF}}{R_{m.}}\right)^2 \\ & + \frac{\sigma^2}{M^2}\sum_{m=1}^{M}\frac{1}{R_{m.}} + \frac{\sigma^2}{R_{m.}}\left(\frac{M-2}{M}\right) \end{aligned} \tag{A8b}$$

$$Var\left(\hat{\beta}_F\right) = \frac{\sigma^2}{\sum_{m=1}^{M} W_m} \tag{A8c}$$

$$\begin{aligned} Var\left(\hat{y}_{msr}\right) = {} & \frac{\sigma^2}{R_{m.}} \\ & + \frac{\sigma^2}{R_{m.}^2 \sum_{m=1}^{M} W_m}\left(R_{m.}^2 - R_{ms}^2 - 2R_{m.}W_m\right) \end{aligned} \tag{A8d}$$

but σ^2 is unknown so is replaced by the estimate s^2 from Equation 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 Equation 8 is

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

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:

$$\hat{\mu} = \frac{1}{\sum_{m=1}^{M} R_{mH}}\sum_{m=1}^{M} R_{mH}\overline{y}_{mH.} \tag{A10a}$$

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

The variances of the estimates are given by:

$$Var\left(\hat{\mu}\right) = \frac{\sigma^2}{\sum_{m=1}^{M} R_{mH}} \tag{A11a}$$

$$Var\left(\hat{\beta}_F\right) = \frac{\sigma^2}{\sum_{m=1}^{M} R_{mH}} + \frac{\sigma^2}{\sum_{m=1}^{M} R_{mF}} \tag{A11b}$$

$$Var\left(\hat{y}_{msr}\right) = \sigma^2 / R_{.s} \tag{A11c}$$

but σ^2 is unknown so is replaced by the estimate s^2 from Equation A3 with $P = 2$ and $\hat{y}_{msr} = \hat{\mu} + \hat{\beta}_s$.

d. F-tests for model-dependent climate response and model-specific discrepancies

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-dependent 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-specific discrepancies ($H_0 : \alpha_m = 0$ 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. t-tests and confidence intervals

The estimates of the expected climate response $\hat{\beta}_F$ in Eqns. A2c, A6c and A10b are linear combinations of the y_{msr}. The y_{msr} are 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 $Var\left(\hat{\beta}_F\right)$. 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

$$\hat{\beta}_F - t_{(100-a/2)\%,N-P}\sqrt{Var\left(\hat{\beta}_F\right)} \;\leq\; \beta_F \;\leq\; \hat{\beta}_F + t_{(100-a/2)\%,N-P}\sqrt{Var\left(\hat{\beta}_F\right)} \quad \text{(A12)}$$

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 $Var\left(\hat{\beta}_F\right)$. The same applies to confidence intervals for the actual values of μ, α_m, γ_{mF} and y_{msr}.

f. Power analysis

1) F-TESTS

Under the alternate hypothesis $H_a : \gamma_{mF} \neq 0$ for any m, then F_γ has a non-central F-distribution with $M - 1$ and $N - 2M$ degrees of freedom and non-centrality parameter λ_γ. Given the expected value of the variance ratio f_γ^2, the Uniformly Minimum Variance Unbiased (UMVU) (Johnson et al. 1995) estimate of λ_γ is

$$\lambda_\gamma = f_\gamma^2\,(N - 2M - 2) - (M - 1) \quad \text{(A13)}$$

The power of the F-test of size $a\%$ for model-dependent climate response is then

$$Pr\left(F_\gamma > F_{(100-a)\%,M-1,N-2M}\;;\;\lambda_\alpha\right) \quad \text{(A14)}$$

Similarly, under the alternate hypothesis $H_a : \alpha_m \neq 0$ for any m, then F_α has a non-central F-distribution with $M-1$ and $N-(M+1)$ degrees of freedom and non-centrality parameter λ_α. Given the expected value of the variance ratio f_α^2, the Uniformly Minimum Variance Unbiased (UMVU) estimate of λ_α is

$$\lambda_\alpha = f_\alpha^2 \left(N-(M+1)-2\right) - (M-1) \tag{A15}$$

The power of the F-test of size $a\%$ for model-dependent climate response is then

$$Pr\left(F_\alpha > F_{(100-a)\%,M-1,N-(M+1)}\ ;\ \lambda_\alpha\right) \tag{A16}$$

663 2) t-TESTS

If the actual value of the expected climate response β_F is β, then T_β may be rewritten as

$$T_\beta = \frac{\hat{\beta}_F}{Var\left(\hat{\beta}_F\right)} = \frac{\sqrt{k}\hat{\beta}_F}{s} = \frac{\sqrt{k}\frac{\hat{\beta}_F-\beta}{\sigma} + \frac{\sqrt{k}\beta}{\sigma}}{\sqrt{\frac{(N-P)s^2}{\sigma^2}\frac{1}{N-P}}} \tag{A17}$$

664 which for $\beta \neq 0$ has the form of a non-central t-distribution with $N-P$ degrees of freedom
665 and non-centrality parameter $\mu = \sqrt{k}\beta/\sigma$, where k is a constant which depends only on on
666 the number of models and runs. Therefore μ may written as $\mu = \sqrt{k}d_\beta$ where d_β is the
667 actual standardised expected climate response.

The power of the t-test of size $a\%$ for non-zero expected climate response is then

$$Pr\left(T_\beta < t_{(a/2)\%,N-P}\ ;\mu\right)\ + \qquad Pr\left(T_\beta > t_{(100-a/2)\%,N-P}\ ;\mu\right) \tag{A18}$$

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 (3)**, 565–573, doi:10.1175/1520-0493(2003)131⟨0565:SOFBAM⟩2.0.CO;2.

Bengtsson, L., K. I. Hodges, and N. Keenlyside, 2009: Will extratropical storms intensify in a warmer climate? *J. Climate*, **22 (9)**, 2276–2301, doi:10.1175/2008JCLI2678.1.

Bengtsson, L., K. I. Hodges, and E. Roeckner, 2006: Storm tracks and climate change. *J. Climate*, **19 (15)**, 3518–3543, doi:10.1175/JCLI3815.1.

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

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 (20)**, 5336–5352, doi:10.1175/2011JCLI4181.1.

Chandler, R. E., 2011: Exploiting strength, discounting weakness: combining information from multiple climate simulators. *Proc. Roy. Soc. A*, **submitted**.

Collins, M., R. E. Chandler, P. M. Cox, J. M. Huthnance, J. Rougier, and D. B. Stephenson, 2012: Quantifying future climate change. *Nat. Clim. Change*, **2 (6)**, 403–409, doi:10.1038/nclimate1414.

Ferro, C. A. T., 2004: Attributing variation in a regional climate change modelling experiment. Tech. rep., EU Project PRUDENCE. URL `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 (10)**, 1141–1158, doi:10.1175/1520-0442(2002)015⟨1141:COAURA⟩2.0.CO;2.

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 (3)**, 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 (11)**, 2573–2585, doi:10.1175/1520-0493(1994)122⟨2573:AGMFTA⟩2.0.CO;2.

Hodges, K. I., 1995: Feature tracking on the unit sphere. *Mon. Wea. Rev.*, **123 (12)**, 3458–3465, doi:10.1175/1520-0493(1995)123⟨3458:FTOTUS⟩2.0.CO;2.

Hodges, K. I., 1996: Spherical nonparametric estimators applied to the UGAMP model integration for AMIP. *Mon. Wea. Rev.*, **124 (12)**, 2914–2932, doi:10.1175/1520-0493(1996)124⟨2914:SNEATT⟩2.0.CO;2.

Hodges, K. I., 1999: Adaptive constraints for feature tracking. *Mon. Wea. Rev.*, **127 (6)**, 1362–1373, doi:10.1175/1520-0493(1999)127⟨1362:ACFFT⟩2.0.CO;2.

Hoskins, B. J. and K. I. Hodges, 2002: New perspectives on the Northern Hemisphere winter storm tracks. *J. Atmos. Sci.*, **59 (6)**, 1041–1061, doi:10.1175/1520-0469(2002)059⟨1041:NPOTNH⟩2.0.CO;2.

Johnson, N. L., S. Kotz, and N. Balakrishnan, 1995: *Continuous Univariate Distributions*, Vol. 2. 2d ed., John Wiley & Sons, Ltd, 752 pp.

714 Kang, E. L. and N. Cressie, 2012: Bayesian hierarchical ANOVA of regional climate-change
715 projections from NARCCAP Phase II. *Int. J. Appl. Earth Obs.*, doi:10.1016/j.jag.2011.
716 12.007.

717 Knutti, R., G. Abramowitz, M. Collins, V. Eyring, P. J. Gleckler, B. Hewitson, and
718 L. Mearns, 2010a: Good practice guidance paper on assessing and combining multi model
719 climate projections. *Meeting report of the Intergovernmental Panel On Climate Change Ex-*
720 *pert Meeting on Assessing and Combining Multiple Model Climate Projections*, T. Stocker,
721 Q. Dahe, G.-K. Plattner, M. Tignor, and P. Midgley, Eds., IPCC Working Group I Tech-
722 nical Support Unit.

723 Knutti, R., R. Furrer, C. Tebaldi, J. Cermak, and G. A. Meehl, 2010b: Challenges in
724 combining projections from multiple climate models. *J. Climate*, **23 (10)**, 2739–2758,
725 doi:10.1175/2009JCLI3361.1.

726 Krzanowski, W. J., 1998: *An Introduction to Statistical Modelling*. John Wiley & Sons, Ltd,
727 264 pp.

728 McDonald, R. E., 2011: Understanding the impact of climate change on Northern
729 Hemisphere extra-tropical cyclones. *Climate Dyn.*, **37 (7-8)**, 1399–1425, doi:10.1007/
730 s00382-010-0916-x.

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

733 Meehl, G. A., C. Covey, K. E. Taylor, T. Delworth, R. J. Stouffer, M. Latif, B. McA-
734 vaney, and J. F. B. Mitchell, 2007: The WCRP CMIP3 multimodel dataset: A new
735 era in climate change research. *Bull. Amer. Meteor. Soc.*, **88 (9)**, 1383–1394, doi:
736 10.1175/BAMS-88-9-1383.

737 Moss, R. H., et al., 2010: The next generation of scenarios for climate change research and
738 assessment. *Nature*, **463 (7282)**, 747–756, doi:10.1038/nature08823.

739 Räisänen, J., 2001: CO2-induced climate change in CMIP2 experiments: Quantification of
740 agreement and role of internal variability. *J. Climate*, **14 (9)**, 2088–2104, doi:10.1175/
741 1520-0442(2001)014⟨2088:CICCIC⟩2.0.CO;2.

742 Rougier, J., M. Goldstein, and L. House, 2012: Second-order exchangeability analysis for
743 multi-model ensembles. *J. Amer. Stat. Assoc.*, **submitted**.

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

747 Solomon, S., D. Qin, M. Manning, Z. Chen, M. Marquis, K. B. Averyt, M. Tignor, and
748 H. L. Miller, (Eds.) , 2007: *Climate Change 2007: The Physical Science Basis*. Cambridge
749 University Press, Cambridge, United Kingdom and New York, NY, USA, 996 pp.

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

752 Stephens, M. A., 1974: EDF statistics for goodness of fit and some comparisons. *J. Amer.*
753 *Stat. Assoc.*, **69 (374)**, 730–737, doi:10.2307/2286009.

754 Taylor, K. E., R. J. Stouffer, and G. A. Meehl, 2012: An overview of CMIP5 and the experi-
755 ment design. *Bull. Amer. Meteor. Soc.*, **93 (4)**, 485–498, doi:10.1175/BAMS-D-11-00094.
756 1.

757 Tebaldi, C., R. L. Smith, D. Nychka, and L. O. Mearns, 2005: Quantifying uncertainty in
758 projections of regional climate change: A Bayesian approach to the analysis of multimodel
759 ensembles. *J. Climate*, **18 (10)**, 1524–1540, doi:10.1175/JCLI3363.1.

760 Yip, S., C. A. T. Ferro, D. B. Stephenson, and E. Hawkins, 2011: A simple, coherent
761 framework for partitioning uncertainty in climate predictions. *J. Climate*, **24 (17)**, 4634–
762 4643, doi:10.1175/2011JCLI4085.1.

763 Zappa, G., L. C. Shaffrey, K. I. Hodges, P. G. Sansom, and D. B. Stephenson, 2012a: A
764 CMIP5 multimodel perspective on the response of North Atlantic and Mediterranean
765 cyclones to climate change. *J. Climate*, **in preparation**.

766 Zappa, G., L. C. Shaffrey, and K. I. Hodges, 2012b: The ability of CMIP5 models to simulate
767 North Atlantic cyclones. *J. Climate*, **in preparation**.

768 Zwiers, F. W., 1987: A potential predictability study conducted with an atmospheric general
769 circulation model. *Mon. Wea. Rev.*, **115 (12)**, 2957–2974, doi:10.1175/1520-0493(1987)
770 115⟨2957:APPSCW⟩2.0.CO;2.

771 Zwiers, F. W., 1996: Interannual variability and predictability in an ensemble of AMIP
772 climate simulations conducted with the CCC GCM2. *Climate Dyn.*, **12 (12)**, 825–847,
773 doi:10.1007/s003820050146.

List of Tables

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

Modelling Center (or Group)	Model Name
Beijing Climate Center, China Meteorological Administration	BCC-CSM1.1
Canadian Centre for Climate Modelling and Analysis	CanESM2
Centre National de Recherches Meteorologiques / Centre Europeen de Recherche et Formation Avancees en Calcul Scientifique	CNRM-CM5
Commonwealth Scientific and Industrial Research Organization in collaboration with Queensland Climate Change Centre of Excellence	CSIRO-Mk3.6.0
EC-EARTH consortium	EC-EARTH
LASG, Institute of Atmospheric Physics, Chinese Academy of Sciences and CESS, Tsinghua University	FGOALS-g2
NOAA Geophysical Fluid Dynamics Laboratory	GFDL-ESM2G GFDL-ESM2M
Met Office Hadley Centre	HadGEM2-CC HadGEM2-ES
Institute for Numerical Mathematics	INM-CM4
Institut Pierre-Simon Laplace	IPSL-CM5A-LR IPSL-CM5A-MR
Atmosphere and Ocean Research Institute (The University of Tokyo), National Institute for Environmental Studies, and Japan Agency for Marine-Earth Science and Technology	MIROC5
Japan Agency for Marine-Earth Science and Technology, Atmosphere and Ocean Research Institute (The University of Tokyo), and National Institute for Environmental Studies	MIROC-ESM MIROC-ESM-CHEM
Max Planck Institute for Meteorology	MPI-ESM-LR
Meteorological Research Institute	MRI-CGCM3
Norwegian Climate Centre	NorESM1-M

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

Model	Runs		Weights					
	Historical	RCP4.5	Two-way		Additive		One-way	
	R_{mH}	R_{mF}	W_{mF}	W_{mH}	W_{mF}	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

List of Figures

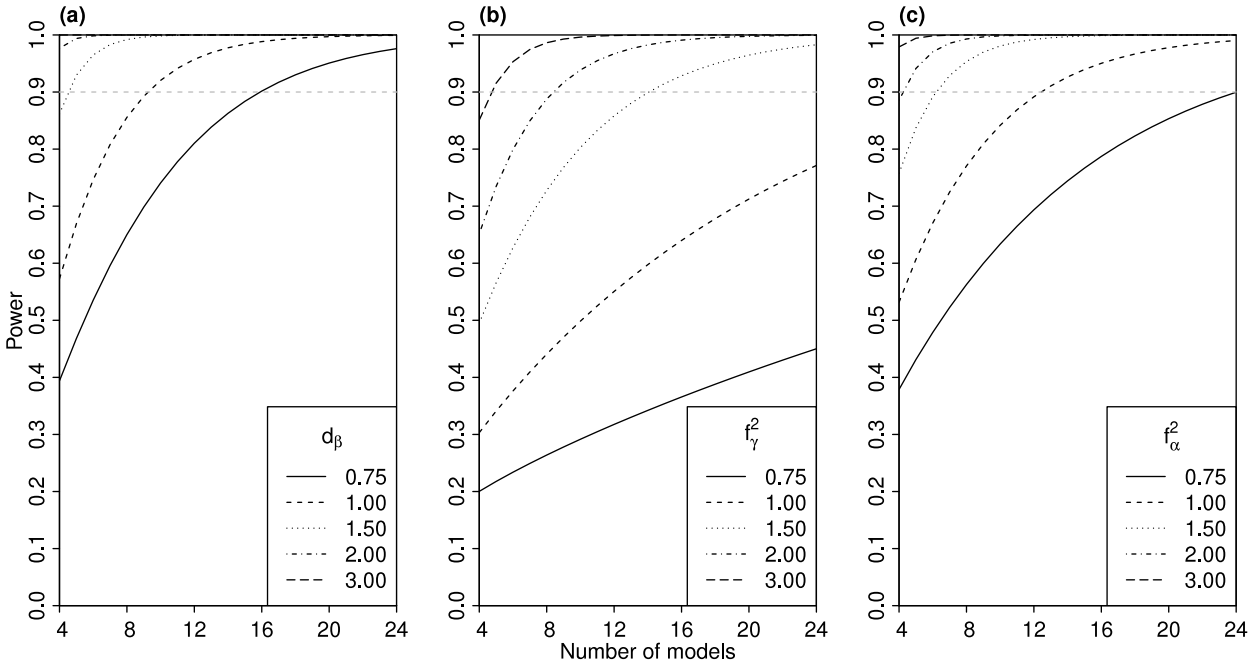

FIG. 1. (a) Power of t-test for significant non-zero expected climate response, (b) power of F-test for significant non-zero model-dependence in the climate response, (c) power of F-test for significant non-zero model-specific discrepancies. Power is calculated for a hypothetical MME which includes a total of four runs from each model, two each of one historical and one future scenario.

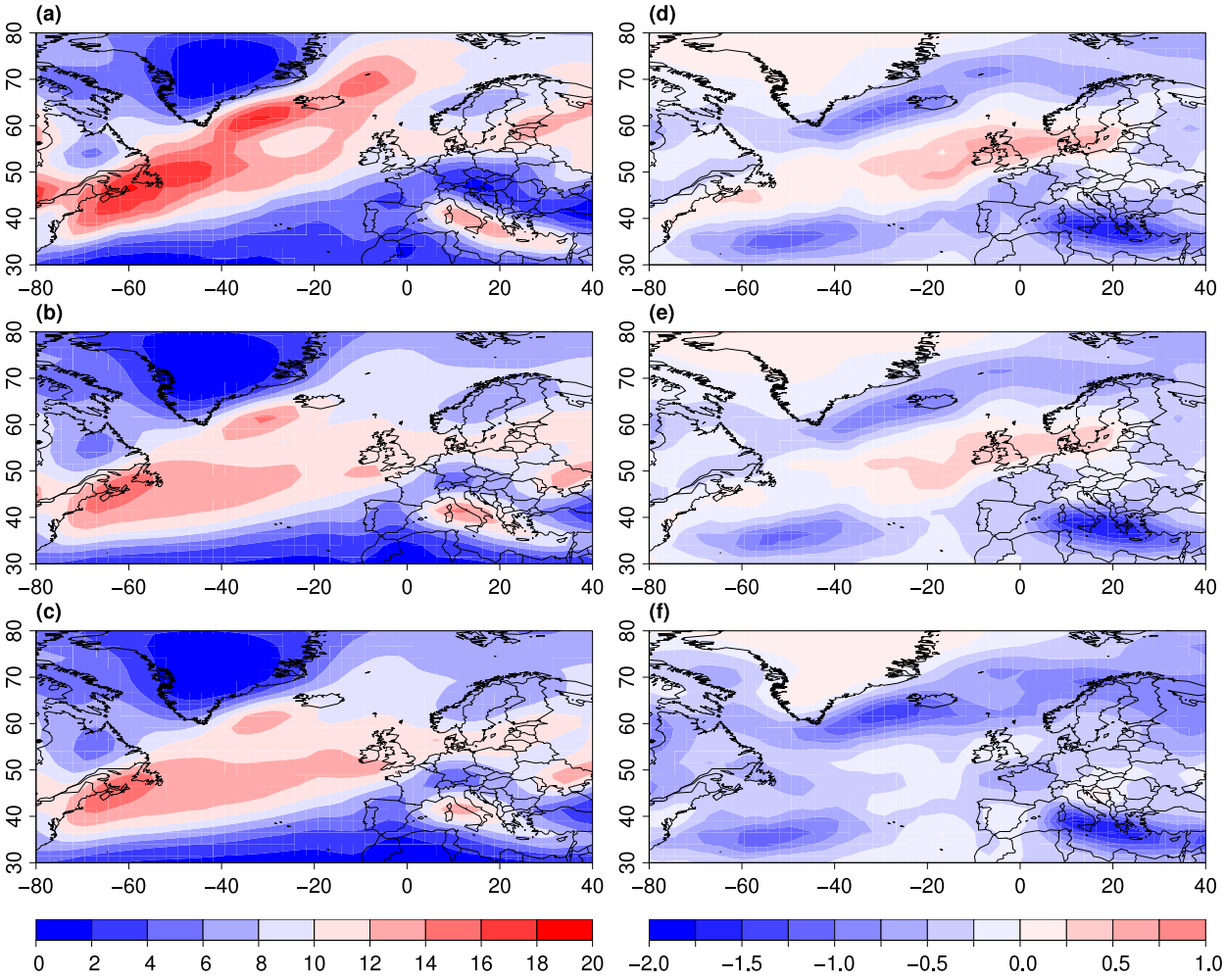

FIG. 2. (a) DJF track density in ERA-Interim, (b) historical DJF unweighted multi-model mean track density simulated by CMIP5 models, (c) RCP4.5 DJF multi-model mean track density simulated by CMIP5 models, (d) unweighted multi-model mean climate response estimate from two-way framework, (e) multi-model mean climate response estimate from the additive framework, (f) multi-model mean climate response estimate from the one-way framework. The CMIP5 models simulate the DJF track density reasonably well. The climate response estimates from the two-way and additive frameworks appear similar. The climate response estimate from the one-way framework fails to capture the projected increase in track density over the UK and Denmark.

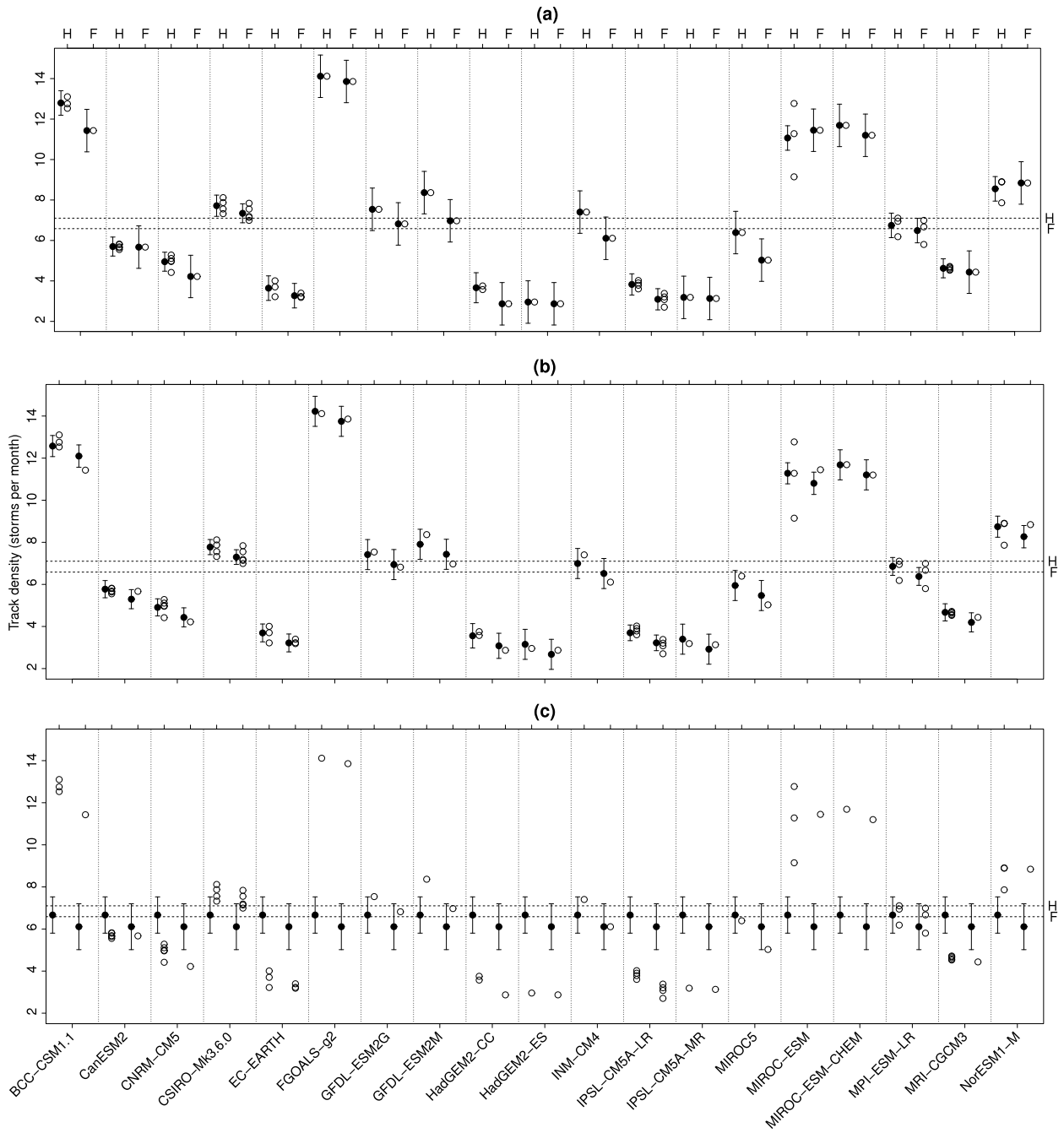

FIG. 3. Estimated mean climates from the three ANOVA frameworks for a grid point (46.5°N 1.25°E) in central France. (a) the two-way framework, (b) the additive framework, (c) the one-way framework. Open points represent individual runs from the historical scenario (H) on the left and the RCP4.5 scenario (F) on the 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. Dashed horizontal lines indicate the usual unweighted multi-model mean estimates of the historical (H) and future (F) climates.

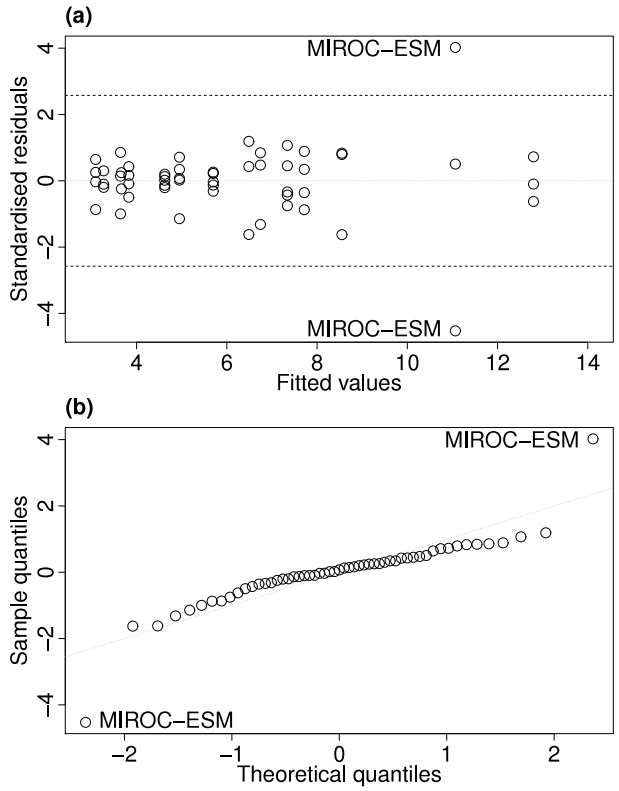

FIG. 4. (a) Plot of standardised residuals against fitted values from the two-way framework. Each point represents one run. Dashed lines indicate the 0.5% and 99.5% quantiles of the standard normal distribution. Realisations appear randomly scattered about zero with the exception of two which are outlying at the 1% level, (b) Quantile-quantile plot of the standardised residuals from the two-way framework. Each point represents one run. Most realisations lie close the expected straight line.

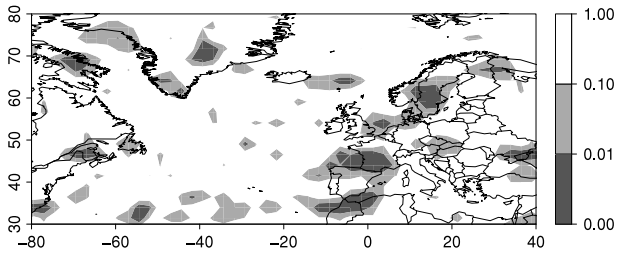

FIG. 5. p-values of the Anderson-Darling test for normality. Small p-values ($p < 0.10$) indicate significant evidence of non-normality. The assumption of normality appears well justified over most of the study region. However several coherent regions of significant non-normality are evident.

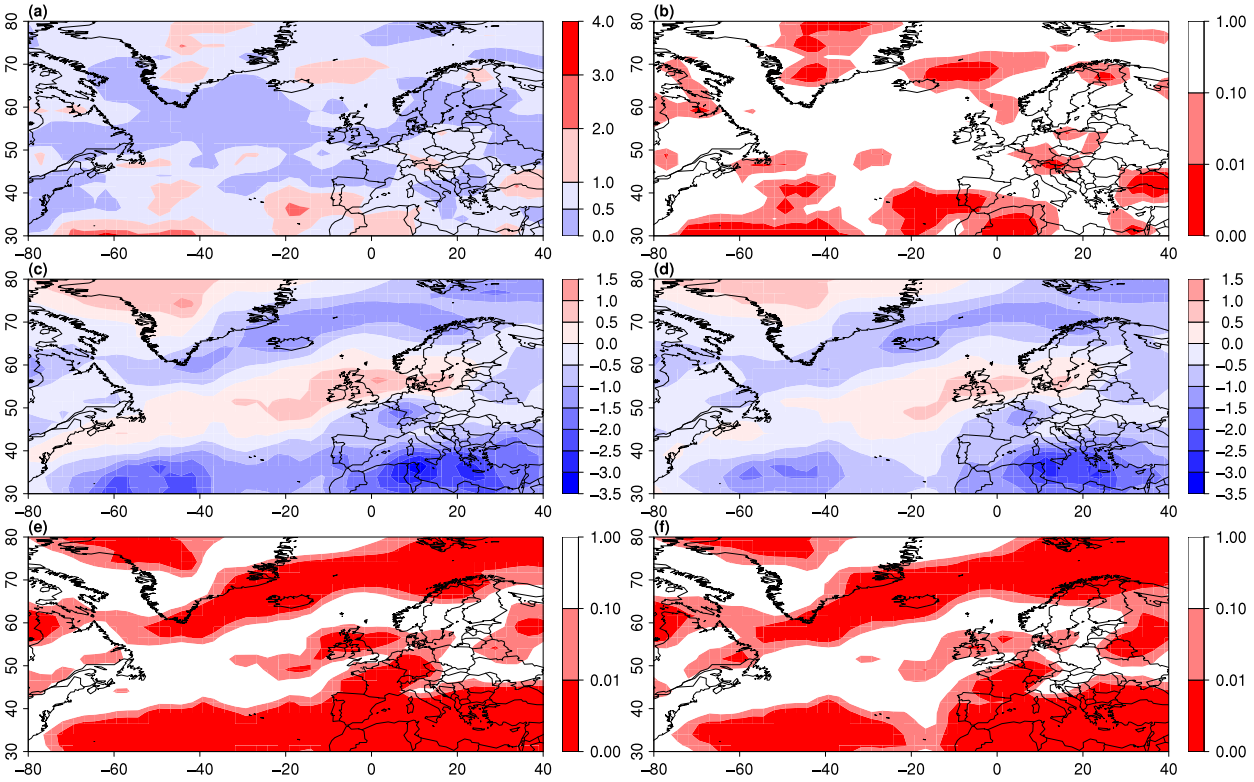

FIG. 6. (a) Variance ratio f_γ^2, (b) p-values of significance test for model-dependent climate response, (c) standardized mean climate response d_β from the two-way framework, (d) standardized mean climate response d_β from the additive framework, (e) p-value of significance test for non-zero mean climate response from the two-way framework, (f) p-value of the significance test for non-zero mean climate response from the additive framework.

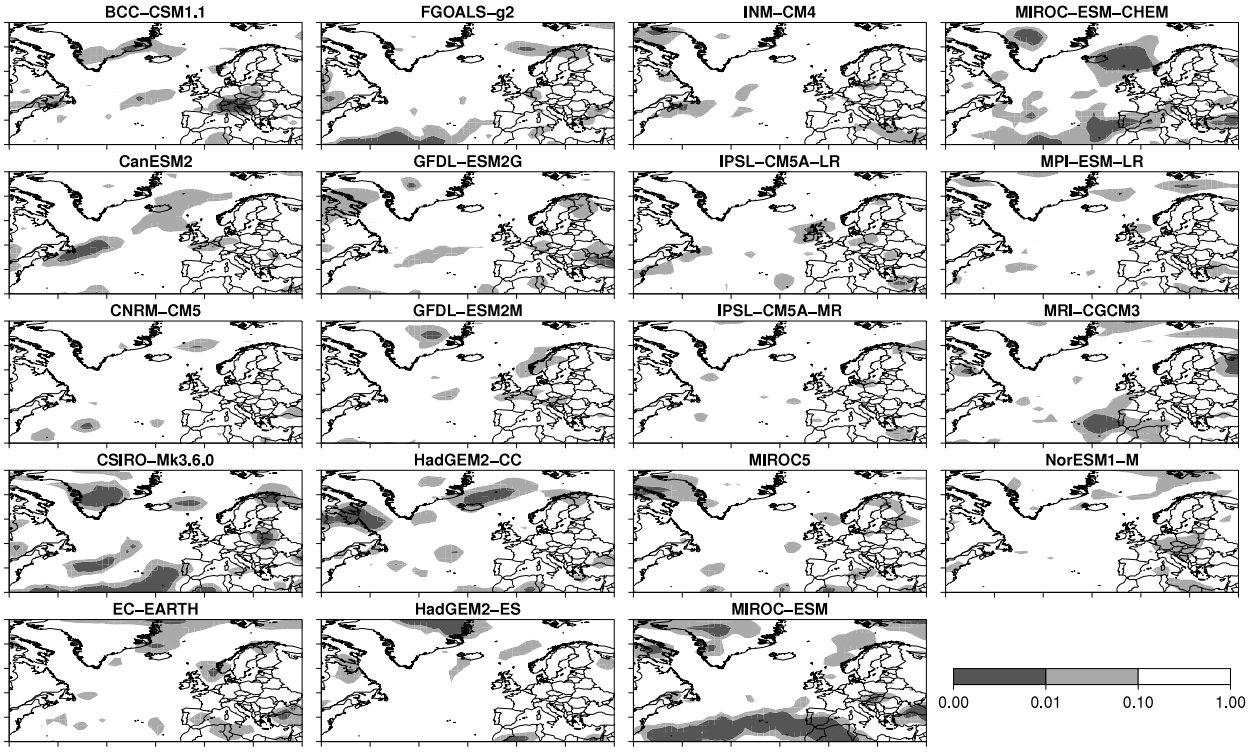

FIG. 7. p-values of the individual t-tests on the γ_{mF} terms. Small p-values ($p < 0.10$) indicate significant evidence that a particular model disagrees with the expected climate response of the MME.

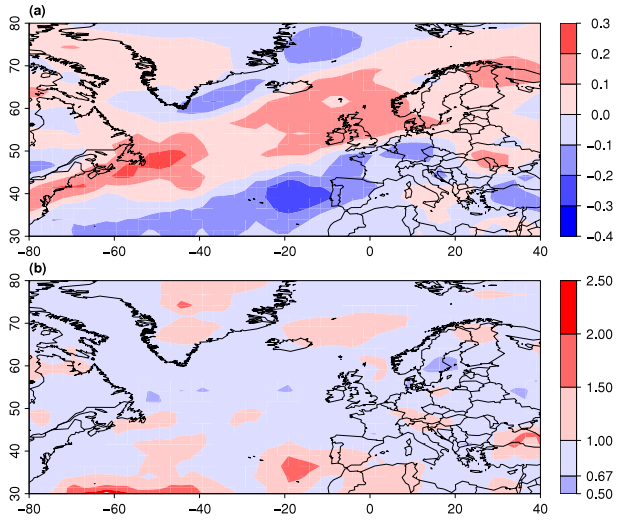

FIG. 8. (a) Difference between expected climate response estimates $\hat{\beta}_F$ from the two-way and additive frameworks, (b) ratio of the estimated uncertainties of the expected climate response from the additive framework and the two-way framework. The two-way framework tends to over estimate the size of the climate response compared to the additive framework. The uncertainty in the additive framework is lower where there is agreement between models on the expected climate response.