An Assessment of the Foundational Assumptions in High-Resolution Climate Projections: The Case of UKCP09

Forthcoming in *Synthese*. DOI 10.1007/s11229-015-0739-8

Roman Frigg
Department of Philosophy, Logic and Scientific Method, LSE
Centre for Philosophy of Natural and Social Science (CPNSS), LSE
Houghton Street
London WC2A 2AE
United Kingdom
Email: r.p.frigg@lse.ac.uk
Tel.: +44 20 7955 7182
Fax.: ++44 20 7955 6845

Leonard A. Smith
Centre for the Analysis of Time Series (CATS), LSE
Pembroke College Oxford
Department of Statistics, University of Chicago
Email: lenny@maths.ox.ac.uk

David A. Stainforth
The Grantham Research Institute on Climate Change and the Environment, LSE
Centre for the Analysis of Time Series (CATS), LSE
Environmental Change Institute, University of Oxford
Department of Physics, University of Warwick
Email: d.a.stainforth@lse.ac.uk

Abstract

The *United Kingdom Climate Impacts Programme's* UKCP09 project makes high-resolution projections of the climate out to 2100 by post-processing the outputs of a large-scale global climate model. The aim of this paper is to describe and analyse the methodology used and then urge some caution. Given the acknowledged systematic, shared errors of all current climate models, treating model outputs as decision-relevant projections can be significantly misleading. In extrapolatory situations, such as projections of future climate change, there is little reason to expect that post-processing of model outputs can correct for the consequences of such errors. This casts doubt on our ability, today, to make trustworthy, high-resolution probabilistic projections out to the end of this century.

Keywords

Climate change; prediction; projection; simulation; model; probability; reliability; emulation; systematic error; decision-making; structural model error.

1. Introduction

There is now a widespread consensus that global warming is real and in large part due to human activities.[1] Simply knowing that the Earth's surface will warm *on the whole* (or *on average*) has value for both mitigation[2] and adaptation strategies, especially when accompanied by other physically understood aspects such as the greater rate of warming of land than of ocean and the warming amplification at higher latitudes. Nevertheless much greater detail is sometimes desired. The question is: to what extent can that desire

[1] The existence of a wide-spread a consensus is documented in (Oreskes 2007); the evidence for the warming being anthropogenic is documented in the most recent IPCC report (Stocker et al. 2013); a shorter summary is (Dessler 2011, Ch. 3).

[2] Knowing even roughly what is likely to happen may be reason enough not to go there.

be met effectively? The impact of climate change on humans (as well as other organisms) occurs at a local scale, and so ideally one would like to know what changes one has to expect in one's immediate environment. For instance, how will the precipitation change in central London by the end of this century? Having reliable answers to such questions would have significant implications for water management, agriculture, health planning, and many other decisions. Robust, reliable answers would aid decision-making (Oreskes et al. 2010; Sexton et al. 2012; Smith and Stern 2011; Tang and Dessai 2012).

The *United Kingdom Climate Impacts Program's* UKCP09[3] project aims to answer exactly such questions by making high-resolution probabilistic projections[4] of the climate out to 2100 based on HadCM3, a global climate model (GCM) developed at the UK Met Office Hadley Centre. The IPCC has confidence that global climate models like HadCM3 have some skill at continental scales and above.[5] This leaves open the question whether decision-relevant high-resolution projections could be constructed with today's models.

The aim of this paper is to describe and analyse the methodology used by UKCP09 and then urge some caution. While this methodology is the only complex model-error-exploring methodology currently deployed to inform decisions with probability projections, UKCP09 is not an isolated phenomenon. In the UK a successor to UKCP09

[3] 'UKCP' stands for United Kingdom Climate Projections and '09' indicates that it was launched for public use in 2009. The project's broad outline is documented in the Briefing Report (Jenkins et al. 2009) (a revised version has been published in 2010); the Science Report (J. Murphy et al. 2010) and two recent papers (Sexton et al. 2012) and (Sexton and Murphy 2012) provide a detailed exposition.

[4] A 'projection' is the 'response of the climate system to emission or concentration scenarios of greenhouse gases and aerosols, or radiative forcing scenarios […]' (Solomon, 2007, 943}. Unlike predictions or forecasts, projections 'depend upon the emission/concentration/radiative forcing scenario used, which are based on assumptions concerning, for example, future socioeconomic and technological developments that may or may not be realised and are therefore subject to substantial uncertainty' (*ibid*.).

[5] 'IPCC' refers to the Intergovernmental Panel on Climate Change, the international body for the assessment of climate change established by the United Nations Environment Programme (UNEP) and the World Meteorological Organization (WMO) in 1988. The Panel's findings are documented in its Assessment Reports. The 4th Assessment report was published in 2007 (Solomon et al. 2007), and the 5th has been released in phases from September 2013 to October 2014.

is being planned, and similar projects are under consideration around the world.[6] The question whether UKCP09 provides decision-relevant projections is widely debated inside the UK; this paper is intended to raise the profile of that discussion, as the answers have implications that reach far beyond the political and scientific context of the UK. UKCP09 has great value as a worked example in the discussion of the strengths and weaknesses of climate simulation in support of good policy worldwide.

Given the acknowledged systematic errors in all current climate models, the fact that many limitations are shared by models, and the fundamental limitations which come into play whenever one extrapolates with imperfect nonlinear models, model outputs cannot be assumed to contain the information necessary to produce reliable probabilistic, multi-decadal projections, particularly on local scales (Smith 2002; Thompson 2013; Frigg et al. 2014). Climate projections are extrapolatory in nature; they rely directly on the information content of the models regarding the interaction of relevant physical processes. If the model simulations are not of high fidelity, then it is questionable whether post-processing of model output, even if involving additional information from observations, would be sufficient to generate trustworthy,[7] high-resolution projections out to the end of this century. This paper elucidates the fundamental assumptions applied in UKCP09, facilitating the questioning of their relevance; this is of particular value because some are widely made in the interpretation of climate modelling experiments.[8]

[6] Similar projects include: Cal Adap (http://cal-adapt.org/precip/decadal/), Climate Wizard (http://www.climatewizard.org/), ClimateimpactsOnline (http://www.climateimpactsonline.com/).

[7] In this paper we shall use the word 'trustworthy' to denote probability forecasts, which one might rationally employ for decision-making purposes using probability theory in the standard way. Such probability forecasts are expected to be robust and reliable, the kind a good Bayesian would make. There may be many justifiable and interesting scientific reasons to construct probability forecasts; our criticism of them in this paper is only in regard to their direct use in decision support (as, for instance, illustrated in the worked examples of UKCP09).

[8] Uncertainty in climate modelling has been given considerable attention, among others, by Parker (2010a, 2013) and Winsberg (2010; 2012). Our discussion has a different focus in that it deals specifically with local climate projection and concentrates on post-processing.

In Section 2 we discuss the aims of UKCP09. In Section 3 we outline the method used to generate high-resolution climate projections. We give considerable space to the description of UKCP09's methods for two reasons. First, even though UKCP09 is widely discussed, its ways and means in generating projections are *terra incognita* outside a narrow circle of experts. We take this paper as an opportunity to make the main outlines of this complex scheme accessible to a wider audience. Second, our criticisms are directed against particular (independent) assumptions of the scheme, and without first introducing these assumptions our discussion would lack a foothold. In Section 4 we discuss the project's handling of structural model error, which is based on what we call the *core assumption* and the *proxy assumption*, and we argue that both assumptions are untenable. In Section 5 we expose general challenges to employing emulators in the climate context, and UKCP09's use of emulators in particular. In Section 6 we discuss the choice of prior probability distributions and in Section 7 we draw attention to issues with initial conditions and downscaling. In Section 8 we consider the issue whether the UKCP09 projections are intended for actual use. In Section 9 we reach the conclusion that the UKCP09's projections should not be regarded as trustworthy projections for quantitative decision support, and propose an initial list of necessary properties for trustworthy projections.

2. UKCP09: Aims and Results

Modelling endeavours can pursue different goals: uncovering mechanisms, understanding causal structures, explaining manifest behaviour, aiding the application of a theory and generating projections are but a few items on a long list. Many of these goals can be (and indeed have been) pursued with climate models. The declared aim and purpose of UKCP09 is to provide decision-relevant projections, on which industry and policy makers can base their future plans. The UKCP09 Briefing Report states:

> 'To adapt effectively, planners and decision-makers need as much good information as possible on how climate will evolve, and supplying this is the aim of the new projections of UK climate change in the 21st century, known as UKCP09. They are one part of a UK government programme of work to put in place a new statutory framework on, and provide practical support for, adaptation.

> The projections have been designed as input to the difficult choices that planners and other decision-makers will need to make, in sectors such as transport, healthcare, water-resources and coastal defences, to ensure that UK is adapting well to the changes in climate that have already begun and are likely to grow in future.' (Jenkins et al. 2009, 9)

In a system as complex as the world's climate, it is absurd to produce a point projection (i.e. a projection saying that a particular event will happen at a particular time with certainty). UKCP09 produces what they dub 'probabilistic projections', which

> 'assign a probability to different possible climate outcomes recognising that […] giving a range of possible climate change outcomes is better, and can help with robust adaptation decisions, but would be of limited use if we could not say which outcomes are more or less likely than others.'(*ibid*., 23)

The challenges many decision makers have to address arise at a local level: flood barriers have to be built in a particular location and to a chosen height, water storage facilities have to be built in suitable locations and so on. For this reason, local user-relevant information about the impacts of climate change is the most useful, assuming of course that it is not mis-informative (Smith and Stern 2011).

UKCP09 tries to meet the demand for decision-relevant information at the local level by producing highly specific information (*ibid*., 6-7). Probabilities are given for events on a 25km grid (which means, for instance, that the projections may differentiate between the impacts of global climate change in London and Oxford, two cities that are only about an hour apart by train). Projections are made for finely defined specific events such as changes in the temperature of the warmest day in summer, the precipitation of the wettest day in winter, or the change in summer-mean cloud amount, with projections blocked into overlapping thirty year segments which extend to 2100. As indicated in the graph on p. 36 of the Briefing Report, it is projected, for instance, that under a medium emission scenario there is a 0.5 'probability level central estimate'[9] for the reduction in summer mean precipitation in central London to be between 20% and 30%.[10] The

[9] We take this phrase to refer to the median of the probability distribution.

[10] The full set of UKCP09 predictions is at http://ukclimateprojections.defra.gov.uk/.

worked examples provided in UKCP09 state explicit design criteria implying this finely defined interpretation is intended.

3. The Architecture of UKCP09

These projections are generated with a method involving both global climate models and elaborate post-processing techniques, where 'post-processing' here refers to operations carried out on model-outputs with the aim of transforming bare simulation results into probabilistic projections. In this section we outline the method with the aim of making its architecture visible and identifying key assumptions. The method can be divided into seven parts: modelling, observation, parameter uncertainty, structural model error, statistical inference, emulation, and downscaling.[11]

Part 1 – Modelling. The cornerstone of UKCP09's exploration of the future of the global climate is HadCM3, which is a coupled atmosphere-ocean general circulation model developed at the Hadley Centre in the UK. The model includes an atmospheric model, a land surface model and an ocean model (which includes a sea ice model) and a coupler.[12] The coupler mediates interactions (such as heat and momentum exchanges) between models. Simulations of processes in the climate system come from nonlinear partial differential equations (PDEs), which define the evolution of continuous fields representing the state of various aspects of the climate system. It is possible neither to integrate PDE's exactly, nor to measure perfectly the continuous fields required to initialise them. In practice equations are discretised (in space and in time). HadCM3's atmospheric component has 19 levels with a resolution of 2.5 degrees of latitude by 3.75 degrees of longitude, which produces a global grid of 96 x 73 grid points. This is equivalent to a surface resolution of about 417 km x 278 km at the Equator, reducing to 295 km x 278 km at 45 degrees of latitude. The oceanic component has 20 levels with a

[11] Our account of the method is based on (J. Murphy et al. 2010, Ch. 3) and (Sexton et al. 2012).

[12] See http://www.metoffice.gov.uk/research/modelling-systems/unified-model/climate-models/hadcm3 (information retrieved on 23 March 2014). Further information about HadCM3 can be found at http://badc.nerc.ac.uk/view/badc.nerc.ac.uk__ATOM__dpt_1162913571289262.

horizontal resolution of 1.25 x 1.25 degrees, and a longer time-step than the atmospheric model.

These models include literally thousands of dynamical variables (the state of each grid point is described by a number of variables such as temperature and pressure). Collectively these variables form the model's state space X. The time evolution of these variables is hoped to mirror the evolution over time of relevant physical variables. These models also include hundreds of parameters (some of them representing physical constants; others defining or controlling small scale processes such as cloud formation which are not resolved explicitly, some purely numerical with no physical counterpart). Parameter values enter, for example, into the transfer of heat, moisture or momentum between the surface and the atmosphere, the reflectivity of sea ice, cloud albedo and behaviour, and convection at sub-grid scales.

To aid the discussion to follow, let us introduce some notation. Let X be the model's state space, and the model's state at time t is $x(t) = \{x_1(t), x_2(t), ...\} \in X$. Let $\alpha = \{\alpha_1, \alpha_2, ...\}$ be the vector of all parameters in the model. The time evolution of HadCM3 is $\varphi_t^C(x;\alpha)$, meaning that (given an initial condition x_0 at time $t = 0$ and a value of α) $\varphi_t^C(x;\alpha)$ specifies the value of the system's dynamical variables $x(t)$ for time $t > 0$ (where t=0 can be in the past or the future). That is, $\varphi_t^C(x;\alpha): X \to X$ maps X onto itself and one can write $x(t) = \varphi_t^C(x_0;\alpha)$. At some places in what follows it is important to emphasise that α assumes a particular value a. In these cases one writes '$\varphi_t^C(x_0;a)$' as shorthand for '$\varphi_t^C(x_0;\alpha = a)$'. If a particular set of values for the parameters are chosen one speaks of a *model version* (D.A. Stainforth et al. 2005). Hence, $\varphi_t^C(x;a)$ is a model version of $\varphi_t^C(x;\alpha)$.

Values of $x(t)$ cannot be computed with pencil and paper methods; a computer is used to numerically calculate $\varphi_t^C(x;\alpha)$. Even today's powerful computers take a long time to

make a run[13] of $\varphi_t^C(x;\alpha)$, and so a less complex model is used for most calculations. To this end an ocean model consisting of a so-called slab model is adopted (i.e. an ocean with no currents and a uniform effective depth of 50m). The role of the oceans in transporting heat is nevertheless represented by an applied atmosphere/ocean heat flux. The result of this manoeuvre is HadSM3, a computationally less demanding model.[14] We write $\phi_t^S(x;\alpha)$ to denote the time evolution of this model, where we take it as understood that the vectors x and α vary with the model structure; they are larger in HadCM3 than in HadSM3 which has fewer variables and fewer parameters.

Part 2 – Observations. UKCP09 provides a method for making projections that incorporates information from observations (Sexton et al. 2012, 2513). It is assumed that there is a vector y representing the climate of the world between 1860 and 2100 (J. Murphy et al. 2010, 51). There is a vector space Y of which y is a member. The vector y is 'a large collection of quantities, where each component is typically indexed by type, and by location and time' (Rougier 2007, 249). It is worth noting that there is no simple relation between x and y (nor between X and Y): the former reflects the model's state at a given time while the latter is a summary of the world's true climate *across* a certain time interval.[15]

The vector y can be decomposed into a historical and a future component (where the convention is adopted that the present is part of the historical component): $y = (y_h, y_f)$.

[13] A 'run' is the calculation of the value of x at some particular future instant of time given a certain x_0 and a set of specific values for α. It is synonymous with the term 'simulation'. With today's climate models a run of a hundred years may take between hours and months depending on the model's complexity and resolution and on the computing hardware utilised.

[14] Going from HadCM3 to HadSM3 roughly doubles the speed of the model.

[15] We note that the notion of a vector representing the world's climate raises many serious questions. Which variables ought to be included? At what time and length scales should its components be defined? And more fundamental, how is climate (as opposed to weather) to be defined in the first place. The documentation of UKCP09 provides little information about how these issues have been resolved. Since nothing in our discussion depends on the definition of y we don't pursue this issue further.

Likewise Y can be separated in historical and a future component: $Y = Y_h \times Y_f$, where ' $\times$ ' denotes the Cartesian product.

Observed quantities are denoted by o, and the values found in actual observations are $\tilde{o}$ (that is, the outcome of an observation is $o = \tilde{o}$); hence $\tilde{o} \in Y_h$.[16] As no observation is perfectly precise, one might consider measurement error. The errors are typically assumed to have a Gaussian distribution. For ease of notation it is assumed that observations and their errors are arranged in vectors of the same structure as the past component of the climate and so one can write:

$$\tilde{o} = y_h + e, \tag{1}$$

where e is the Gaussian error distribution.[17]

Part 3 – Parameter Uncertainty. One problem in determining the future values of x is that 'the available information is seldom precise enough to allow the appropriate value of a given parameter to be accurately known' (J. Murphy et al. 2010, 37).[18] Not knowing what value of α to use in calculations, yet assuming there is one, 'gives rise to the parameter component of model error' (*ibid*.).[19]

[16] Given our provisos about the definition of y, selecting observations as indicators of the climate vector is an equally difficult task. UKCP09 use so-called 'pseudo-observations': 'We obtain these by using two or three alternative data sets for each observed quantity, from which we generate 100 pseudo-observations made by adding random linear combinations (where coefficients sum to one) of the different data sources […] regridded onto the HadSM3 grid' (Sexton et al. 2012, 2517). Again, nothing in the discussion to follow depends on how exactly observations are treated and so we set this issue aside.

[17] Of course these distributional assumptions are often questionable; for example they cannot hold for precipitation which is positive definite.

[18] Arguably parameters of an imperfect model are not uncertain but rather indeterminate, as there is no ideal set of parameter values which will make the model adequate for all predictive purposes, as would be the case if the model structure was perfect and the values of the parameters were well defined but simply unknown (Du and Smith 2012).

[19] This assumption is controversial. (Smith 2006) argues that for imperfect models appropriate values (leading to trustworthy forecasts) may not exist. For want of space we set these worries aside and

The technique of a *perturbed parameter ensemble* (PPE) is designed to address this difficulty (Allen and Stainforth 2002).[20] The leading idea of a PPE is to calculate the future values of x for a number of different values of α, where these values ideally are chosen in a manner that sample the diversity of reasonable values. If, for instance, we are uncertain about the 'best' value of parameter α_2 but believe that it lies between $a_{2,\min}$ and $a_{2,\max}$, then one carries out calculations of x for many values in the interval $[a_{2,\min}, a_{2,\max}]$.[21] The variability of the outcomes provides a sense of the *sensitivity* of the model. Calculating future values of x for a number of different parameter values amounts to constructing a PPE because the variation of the parameter values amounts to perturbing the parameter yet without changing the mathematical structure of the model (because the formulation of the equations remain unchanged).

Given the complexity of a model like HadCM3 relative to state-of-the-art computational capacity only a relatively small number of runs are available. The question then is how to construct a PPE for a model with 100s of parameters if one only has a small number of runs. UKCP09 first limits this problem by restricting attention to predominantly atmospheric parameters and then solicits parametrisation experts to identify those parameters which they believe control the crucial processes in the system, on which the future values of x depend sensitively; these experts are then asked to specify plausible intervals for these[22] parameters (J. Murphy et al. 2010, 37, 49). This process led to the identification of 31 crucial parameters and the definition of their associated plausible intervals.

Restricting attention to these 31 parameters, one can ask: what is the variation of future values of x *given* the diversity in α ? UKCP09 quantify uncertainty in x

proceed as if the question was one of uncertainty not indeterminacy; for more on this point see (Smith and Stern 2011).

[20] We note in passing the lack of unanimity on whether the second 'P' of PPE stands for 'parameter', 'parameterization', or 'physics'.

[21] For a discussion of what 'best' might mean see also (Parker 2010b).

[22] Note the difference between the range of reasonable model-parameter values within the model and the uncertainty in the value of the corresponding physical parameter, when such a thing exists.

probabilistically, meaning that they specify a distribution over the uncertainty interval associated with α. The choice made is the following:

> 'we used trapezoidal distributions for the continuous parameters whereby the middle 75% of the expert range was considered equiprobable, and then probability density reduced linearly to zero at the extreme values. For the discrete parameters, the different levels were considered equiprobable. Parameters are assumed independent so that it is straightforward to determine probabilities for every possible combinations [sic] of parameter values.' (Sexton et al. 2012, 2517)

We write $T(\alpha)$ to refer to this distribution.

Part 4 – Model Error. UKCP09 uses what they call the 'best input assumption': 'that for a given climate model there exists a best set of model parameters, [a^*], that provide the best simulation of the true climate, y' (Sexton et al. 2012, 2521). They add that 'due to imperfections in the climate model, even the best input has deficiencies in simulating the true climate' (*ibid.*). The response to this point is to introduce the so-called discrepancy term d. This term is defined as the distance between the best model simulation and the true climate.[23]

A form of the discrepancy term is determined as follows. Use the climate model $\varphi_t^S(x_0;\alpha)$ to calculate the components of a vector that has the same structure as y. Call this vector $\Phi(\alpha)$, where the explicit mention of α indicates its dependence on the parameter α; obviously $\Phi(\alpha) \in Y$. The vector for the best input assumption is $\Phi(a^*)$. Assuming some metric on Y, the discrepancy is the difference between y and $\Phi(a^*)$. On the assumption that it is well-defined, the value of the discrepancy is uncertain (meaning imprecisely known, not indeterminate) and so should be expressed as a probability distribution ε. So one can write (Sexton et al. 2012, 2521):[24]

[23] There are a host of challenges here, as climate is a distribution, the model climate and the real climate are not in the same state space, whatever notion of 'best' is taken the simulation model will not be 'best' for all target variants, and so on. These points raise important questions about how the distance is measured and what the discrepancy is intended to represent in practice in the climate case.

[24] UKCP09 expresses this by writing (in our notion) $y = \varphi(a^*) + \varepsilon$; see equation (1) in (Sexton et al. 2012, 2521). However this is not a correct formal expression of the concept of a discrepancy term

$$y = \Phi(a^*) + \varepsilon \tag{2}$$

As above, this equation can be decomposed into a historical and a future part:

$$\begin{aligned} y_h &= \Phi_h(a^*) + \varepsilon_h \\ y_f &= \Phi_f(a^*) + \varepsilon_f \end{aligned} \tag{4}$$

The discrepancy term therefore

> 'effectively represents how informative the climate model is about the *true climate*, and it measures the difference between the climate model and the *real climate* that cannot be resolved by varying the model parameters. Such differences could arise from processes which are entirely missing from the climate model, or from fundamental deficiencies in the representation of processes which are included, through (say) limited resolution or the adoption of an erroneous assumption in the parameterisation scheme.' (Sexton et al. 2012, 2515, emphasis added)[25]

In brief, by adding the discrepancy term to the model one can glean 'what the model output would be if all the inadequacies in the climate model were removed, without prior knowledge of the observed outcome' (Sexton et al. 2012, 2521).

The use of the discrepancy term is based on two assumptions. The first assumption is 'that the climate model *is* informative about the real system' (*ibid*., original emphasis). This amounts to saying that at least for the best input parameter a^* the model output is close to the real climate: '[a^*] is not just a 'statistical parameter', devoid of meaning: it derives its meaning from the physics in the climate model being approximately the same as the physics in the climate' (Rougier 2007, 253). We call this the *informativeness assumption*.

---

because, as noted above, y and x are not members of the same vector space and hence cannot be added. For a discussion of 'subtractability', see (Smith 2006) and references therein.

[25] See also (J. Murphy et al. 2010, 63-64).

The second assumption concerns the distribution ε and says that this distribution is Gaussian. We call this the *Gaussianity assumption*. It is then also assumed that ε and e and α are 'probabilistically independent' (*ibid*.).

Not being omniscient, one cannot just make a comparison between model outputs and reality. A crucial leap UKCP09 makes is to use a multi model ensemble (MME) as a proxy for the truth:

> 'Our key assumption is that sampling the effects of structural differences between the model chosen for the PPE and alternative models provides a reasonable proxy for the effects of structural errors in the chosen model relative to the real world.' (Sexton et al. 2012, 2516)

> 'this approach is based on the assumption that structural differences between HadSM3 and other models are a plausible proxy for the uncertain effects of structural errors in how HadSM3 represents climate processes in the real world.' (Sexton et al. 2012, 2526)

> 'It is based on the judgement that the effects of structural differences between models can be assumed to provide reasonable *a priori* estimates of possible structural differences between HadSM3 and the real world. We take a given multi-model ensemble member as a proxy for the true climate, and use our emulator of HadSM3 to locate a point in the HadSM3 parameter space which achieves the best multivariate fit between HadSM3 and the multi-model member' (J. Murphy et al. 2010, 64)

The MME in question contains 12 models (see (Sexton et al. 2012, 2519) for details). The view expressed in these quotations is that measuring the average distance of HadSM3 to a set of different models yields a similar result as measuring its distance to the real world. We call the view that a MME is a trustworthy proxy for the real world the *proxy assumption*.

So the aim is to determine the parameters of the distribution ε by comparison of the outputs of HadSM3 with the outputs of other models in an MME. The leading idea behind the actual calculations is to first determine the best HadSM3 analogue for each model in the ensemble. Having found the best analogue, one can calculate the error b, essentially the difference between the two model outputs. The procedure is repeated for

each MME member, giving 12 b's. From these the mean and the covariance matrix of ε are determined.[26]

Under the proxy assumption, this procedure quantifies the additional uncertainty due to structural model error. One can then add this uncertainty to the uncertainty about values of y obtained in Part 3 and thereby obtain the total uncertainty, which now includes an estimate of the uncertainty due to structural model error. As noted by Murphy *et al*. (2007) it is important to stress that this is a lower bound.

Part 5 – Statistical Inference. UKCP09 provides probabilistic projections. With the above in place one can now say how statistical inferences are drawn from the information gathered in Parts 1-4. The aim is to calculate $p(\hat{y}_f \mid \tilde{o})$, the probability of a particular future climate $\hat{y}_f \in Y_f$ given past observations. We use '$p(\hat{y}_f \mid \tilde{o})$' as a shorthand for '$p(\hat{y}_f \mid o = \tilde{o})$'; and the same convention is used below for $p(\hat{y}_f \mid \tilde{o}, a)$, $p(a \mid \tilde{o})$ and $p(a)$.

On the assumption that the discrepancy term compensates for any difference between model outputs and the true climate, the only residual uncertainty is parameter uncertainty. As per Part 3, this uncertainty is assumed to be understood and quantified – by the trapezoidal prior distribution $T(\alpha)$ – and so one can use the law of total probability to calculate the value of $p(y_f \mid o = \tilde{o})$ as a function of the parameter uncertainty:

$$p(\hat{y}_f \mid \tilde{o}) = \int_A p(\hat{y}_f \mid \tilde{o}, \alpha)\, p(\alpha \mid \tilde{o})\, d\alpha\,, \qquad (5)$$

where A is the 31-dimensional uncertainty interval $[a_{1,\min}, a_{1,\max}] \times ... \times [a_{31,\min}, a_{31,\max}]$.

Applying Bayes' theorem to $p(\alpha \mid \tilde{o})$ yields the posterior distribution:

[26] For details see (Sexton et al. 2012, 2525-27).

$$p(\hat{y}_f \mid \tilde{o}) = \frac{1}{p(\tilde{o})} \int_A p(\hat{y}_f \mid \tilde{o}, \alpha)\, p(\tilde{o} \mid \alpha)\, p(\alpha) d\alpha, \qquad (6)$$

where $p(o = \tilde{o})$ can again be expanded using the law of total probability:

$$p(\tilde{o}) = \int_A p(\tilde{o} \mid \alpha) p(\alpha) d\alpha. \qquad (7)$$

This is the core equation of UKCP09.[27]

Let us have a look at the terms in the equation. The last term in the integral is the trapezoid distribution over the uncertainty intervals: $p(\alpha) = T(\alpha)$. The middle term is the *likelihood function*. It evaluates how likely the actual observations are in light of the climate model and the discrepancy term. Recall that $y_h = \Phi_h(a^*) + \varepsilon_h$. Generalising, let $y_h(\alpha)$ be the climate retrodicted by the model for parameter α. Trivially we have $y_h(\alpha) = \Phi_h(\alpha) + \varepsilon_h$. Furthermore recall $o = y_h + e$. Hence

$$o = \Phi_h(\alpha) + \varepsilon_h + e. \qquad (8)$$

Finally recall further that ε_h and e are distributions. Hence this equation provides probability distribution for the probability of past observations conditional on the parameter α. This gives a probabilistic weight to the actual observation $\tilde{o}$, and this weight is the probability $p(\tilde{o} \mid \alpha)$ (also known as the likelihood function).

The first term in the integral can be dealt with in the same way. $y_f = \Phi_f(a^*) + \varepsilon_f$ induces a probability distribution $\Phi_f(\alpha) + \varepsilon_f$, which depends on α, over future climates and this distribution can be used to give a probabilistic weight to the future climate under consideration, $\hat{y}_f$. This is probability $p(\hat{y}_f \mid \tilde{o}, \alpha)$.

[27] This is equation (5) in (Sexton et al. 2012, 2523); for a discussion of the derivation see (Rougier 2007).

The above integration is over the entire interval A. Such an integration can be carried out only if both $\Phi_h(\alpha)$ and $\Phi_f(\alpha)$ are known for all values of α in A. This is not the case. In fact, to explore the uncertainty of future values of x brought about by the variation in these 31 parameters, 280 runs were made with HadSM3 (the simplified model). Later 17 runs with HadCM3 (the larger model) were added and information from the two combined.

Part 6 – Emulator. Emulation is a powerful tool which aids understanding the behaviour of complex computer models when the total number of runs is constrained by technology. The small number of runs underlying UKCP09 does not approach a complete exploration of the parameter space. Indeed it is too small to provide an understanding of the diversity of outcomes. And to carry out the above integrations one would have to know how $\Phi_h(\alpha)$ and $\Phi_f(\alpha)$ depend on *all* parameter values and not only the ones used in the actual model runs. Filling the gaps between any finite number of model runs in a large parameter space is the task performed by a statistical tool called an *emulator*.[28]

The emulator predicts the output x for any parameter value at any time t. For a fixed initial condition and a particular future time t, $\varphi_t^S(x;\alpha)$ specifies a functional relationship between α (which now is the independent variable) and x at the time of interest (now the dependent variable). This relationship defines a surface in X, so $\varphi_t^S(x;\alpha)$ is referred to as the *target surface*. Assuming that x is a sufficiently smooth function of α, an emulator is built giving values of x for *all* α. The emulator does not attempt to recreate the internal dynamics of the model but rather, builds up a distribution for each outcome solely on the basis of the data points obtained in simulations and statistical assumptions. An emulator is akin to a statistically satisfying curve-fitting algorithm; it is more than mere curve-fitting in that provides a distribution from which 'curves' (the mean, the median) can be extracted complete with local accuracy estimates. Let us denote the surface over output variables that the emulator

[28] Note that emulators may elsewhere be called 'surrogate models', 'meta-models', and 'models of models'. A general introduction to emulation (unrelated to UKCP09) can be found at http://mucm.aston.ac.uk/MUCM/MUCMToolkit/index.php?page=MetaFirstExample.html.

provides by $\psi_t(x;\alpha)$. For the known points at which a climate model run is available one must have $\psi_t(x;\alpha)=\varphi_t^S(x;\alpha)$. As a statistical tool the emulator does not provide an exact curve $\psi_t(x;\alpha)$ because the exact location of the curve $\psi_t(x;\alpha)$ is uncertain (except at the simulated points). Rather the emulator provides a distribution of the location of $\psi_t(x;\alpha)$ in X. We call this distribution $\Psi(x;\alpha)$.

What we have just sketched is what we call a *complete emulator*. It is complete in the sense that it emulates the complete set $x(t)=\{x_1(t),x_2(t),...\}$ of dynamical variables. On the basis of the values of $\{x_1(t),x_2(t),...\}$ one can then compute the various climate variables we are interested in, for instance global mean temperature, daily mean precipitation, etc. This is because climate variables of interest are functions of the complete set of dynamical variables (complete relative to the model, that is). Let *v* be a climate variable of interest, then we can write: $v(t)=f(x_1(t),x_2(t),...)$, where the specifics of *f* depend on *v* and $\{x_1(t),x_2(t),...\}$. If, for instance, $\{x_1(t),x_2(t),...\}$ contains temperatures and if *v* is GMT, then *f* is simply a weighted average.

A complete emulator is expensive to build. So in practice what is being built is what we call a *targeted emulator*. A targeted emulator is one that emulates a particular climate variable *v* (e.g. GMT) *directly* rather than first emulating $\{x_1(t),x_2(t),...\}$ and *then* calculating GMT from the $\{x_1(t),x_2(t),...\}$ using *f*. So a targeted emulator takes as input the parameters $\alpha=\{\alpha_1,\alpha_2,...\}$ and provides as output the variable of interest, e.g. GMT, as a function of $\alpha=\{\alpha_1,\alpha_2,...\}$ and time *t*. Such an emulator provides a distribution $\Psi(v;\alpha)$. If *v* is a scalar then the emulator is a *scalar emulator*; if *v* is a vector, it is a *multivariate emulator*. UKCP09 uses a multivariate emulator that emulates a vector with 12 components, where the components are climatic variables such as the sea surface temperature, precipitation, etc. (Sexton et al. 2012, 2518).[29]

[29] UKCP09 does so indirectly in the sense that it emulates the coefficients of a set of basis vectors for the output space in question. Nothing in the discussion that follow depends on this.

The crucial point is that the emulator replaces the model in the process of statistical inference: 'Put simply, the ensemble of model evaluations is used to build the emulator, and then the emulator is used in the inference' (Sexton et al. 2012, 2523); see also (Rougier 2008, 827.829). This means that in actual calculations the vectors $\Phi_h(\alpha)$ and $\Phi_f(\alpha)$, which determine the probabilities $p(\tilde{o}\,|\,\alpha)$ and $p(\hat{y}_f\,|\,\tilde{o},\alpha)$ in the above integral are calculated not with the model itself but with the emulator instead.

Taking the results from Parts 5 and 6 together we gain an important insight: in actual calculations the probabilities offered as projections are computed on the basis of three items: emulator probabilities $\Psi(v;\alpha)$, the discrepancy distribution ε, and the distribution $T(\alpha)$, along with the observations $\tilde{o}$ and their uncertainties e.

Part 7 – Downscaling. The model calculations are done on the HadSM3 grid, which has a resolution of approximately 300km. UKCP09, however, provides projections on a much finer scale of 25km. In order to generate projections at that level of detail, a downscaling method is introduced to derive such information from the global model simulations done in Part 1: 'Finally, to make the projections suitable for impacts and adaptation assessments, we use a further ensemble of the Met Office regional climate model (HadRM3) to downscale the projections from the global Met Office model to a resolution of 25 km' (J. Murphy et al. 2010, 40). Dynamic downscaling of this sort involves running a high-resolution limited area regional model over the geographical domain of interest, with the boundary conditions provided by a global model. In UKCP09 this final stage involves an ensemble using HadCM3 (Murphy et al. 2010). A scaling process (referred to as timescaling) is also used to derive time dependent information through the 21st century because the HadSM3 simulations generate only the equilibrium response to doubling atmospheric carbon dioxide (Murphy et al. 2010). This downscaling and timescaling are critical steps in the process which enables UKCP09 to offer time-dependent information at local scales. The details, however, are beyond the scope of this paper and are not critical to the other issues discussed below.

The endeavours of these seven parts taken together produce the projections we have seen in the last section. We now turn to a discussion of these projections. We discuss each of the crucial ingredients in turn.

4. Structural Model Error

Accounting for structural model error is a critical step in any application for decision support. The discrepancy term and the use of emulation provide important new tools to this end. The effective and appropriate use of these tools hinges on the informativeness assumption and the proxy assumption. We do not dispute the deep importance of these new tools in general, nor the need to improve traditional approaches to model error, but in this section and the one that follows we discuss the efficacy of these tools as deployed in UKCP09.

4.1 The Discrepancy Term and Structural Model Error

Let us start by explaining in more detail the challenge that leads to the introduction of the discrepancy term. The issue is *structural model error* (SME) (Beven 2012; Kennedy and O'Hagan 2001; McWilliams 2007; Smith 2002). Like every model, HadCM3 has its imperfections. That is to say that in specifying $\varphi_t^C(x;\alpha)$, a number of strongly idealising assumptions are made in terms of the relationship between model and reality. There are known flaws forced upon us by technological limitations. These include distortion of the topography of the earth: mountain ranges like the Andes are systematically too smooth and too short, small volcanic islands chains, some easily visible in satelite photographs due to the effect they have on atmospheric circulation via cloud tracks, do not exist in the model, and of course cloud fields themselves cannot be simulated realistically at the available resolution. Recall that most model runs on which UKCP09's projections are based are done with HadSM3, which has an ocean with no dynamically resolved currents and a uniform depth of 50m everywhere. Solutions of the discretized PDE differ from those of the original PDE, and the PDE itself differs from the true equations of the world (assuming such equations exist at all).[30]

[30] Furthermore, there are limitations to our scientific understanding of the climate system and there may be relevant factors and process that we are simply unaware of – there may be unknown unknowns which would lead us to alter the equations of the model even under our current computational constraints.

Inasmuch as SME is due to shortcomings in the equations of the model, the challenges it poses to producing projections cannot be resolved by varying the model's parameters. If a model has SME this means that the time evolution of an ensemble will, eventually, differ from that of a better model and indeed reality itself (if a relevant distribution can be associated with reality). Because this difference is due to the mathematical structure of the model equations no adjustment of the parameters in a model of that structure can remove this difference (Smith 2002). The crucial question is: how soon do nontrivial effects of SME manifest themselves in a given situation? And to what extent can a model with SME still be informative about the target system? On what timescales does the science underlying the model suggest that a decision maker is likely to encounter a big surprise if the model outputs are taken as trustworthy?

UKCP09 proposes the discrepancy term as a solution to SME. The message is that the uncertainties due to SME can be estimated and taken into account in projections. Adding this term to actual model runs is presented as giving us 'what the model output would be if all the inadequacies in the climate model were removed, without prior knowledge of the observed outcome' (Sexton et al. 2012, 2521), and the UKCP09 science report calls the discrepancy term 'an appropriate means of quantifying uncertainties in projected future changes' (J. Murphy et al. 2010, 66). In this section we consider whether its use for the provision of quantitative decision support is justified. For the sake of argument we grant the assumption that the discrepancy is Gaussian and focus on the informativeness and proxy assumptions.

4.2 The Informativeness Assumption

Recall that the informativeness assumption says that the distance between the model and the truth is relatively small; so small indeed that 'the physics in the climate model being [sic] approximately the same as the physics in the climate.' (Rougier 2007, 253). How good is this assumption?

That systematic errors in the models in question lead to non-trivial macroscopic errors of simulation, of the past and of the future, is not disputed (James M. Murphy et al.

2007). In nonlinear models such as HadCM3 even extremely small SME can result in there being a significant discrepancy between model and target (Judd and Smith 2004; McWilliams 2007; Frigg et al. 2014). Seager *et al.* (2008) have noted their inability to reproduce a realistic dust bowl of the 1930's even given the observed sea-surface temperatures. This is not a negligible shortcoming when one is focused on the resolution offered by UKCP09. Given these systematic errors there are lead times at which the failure of the model to simulate realistic weather[31] results in differences in feedbacks, which cause the climate of the model to differ from that of the planet. When the models used are not close to the target, simple linear transformations are inappropriate and the informativeness assumption fails. The informativeness assumption is a crucial building block in the edifice of UKCP09 and if that assumption fails, then the accuracy and decision-relevance of projections is called into question.[32]

The same conclusion can be reached from a different angle. Figure 1 shows the various model global mean temperatures (GMT) over the period 1900 to 2000 for the CMIP5 ensemble (Meehl et al. 2009), the MME which is the updated version of CMIP3 used by UKCP09 (Sexton et al. 2012, 2519) – the figure for CMIP3 is very similar. Note that while all models show warming between 1900 and 2000, their average temperatures differ by as much as 3 degrees and the details of their temperature curves vary tremendously. Aggregate variables such as GMT 'smooth out' local variation so, of course, different local behaviours can give rise to the same average behaviour. That the models' GMTs differ significantly, however, is an indication that their local climates are also significantly different. Local features such as sea ice, snow coverage and surface vegetation will vary and thus local feedbacks in the models will be substantially different from each other, and from those in reality. The magnitude of the range of global mean temperature hindcasts of the last century, casts significant doubt on the viability of the informativeness assumption for 25 km projections to the end of this century.

[31] Even today's best climate models do not simulate blocking realistically. For a discussion of this point see (Smith and Stern 2011) and Hoskins' review of UKCP09 (available at http://ukclimateprojections.metoffice.gov.uk/23173).

[32] The question of what good science should report on the lead times beyond those on which quantitative guidance is informative is a separate issue.

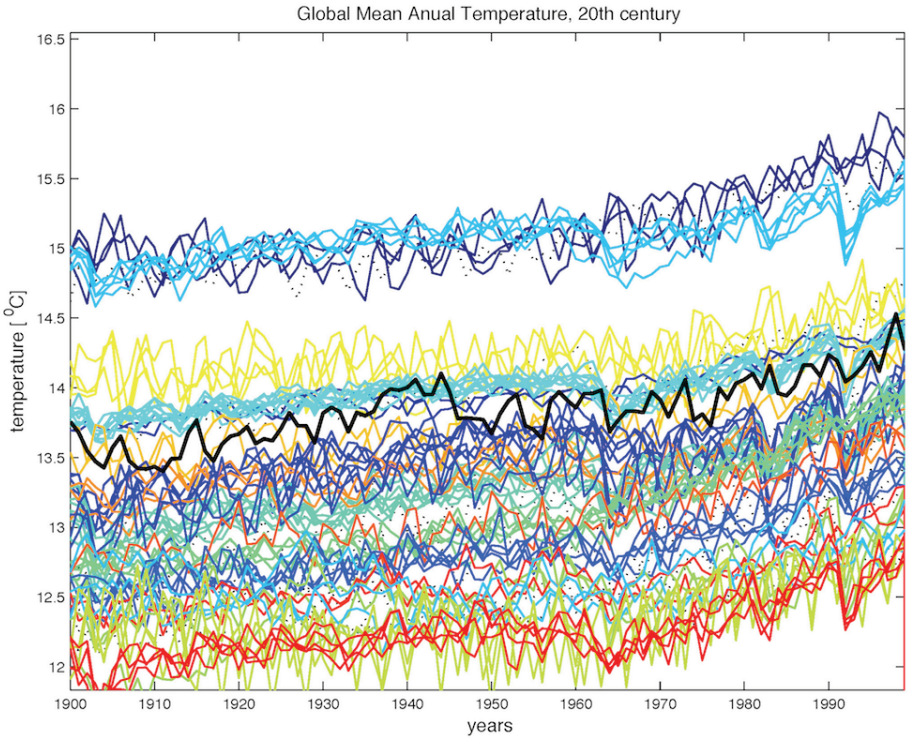

Figure 1: Model global mean temperatures over the period 1900 to 2000 for the CMIP5 ensemble.

There is also a statistical argument based on results due to Murphy *et al.* (2004, 2007), which is clear on the limitations noted here, casting doubt on the informativeness assumption. Murphy *et al*. consider 32 climate variables $v^{(i)}$, $i = 1,...,32$. For each of these variables there is a time series of past observations. These form a vector $m^{(i)}$ with the components $m_j^{(i)}$, where j ranges over the available data points (to keep notation simple we assume that there are n data points for all 32 variables; nothing here depends on that). One can then calculate the mean and the variance of each time series: $\mu^{(i)} = (1/n)\Sigma_{j=1}^{n} m_j^{(i)}$ and $\sigma^{(i)} = (1/n)\Sigma_{j=1}^{n}(m_j^{(i)} - \mu^{(i)})^2$. Assuming that the error is Gaussian, these two parameters define a Gaussian $G^{(i)}(v^{(i)}) = c\exp[(v^{(i)} - \mu^{(i)})^2 / 2\sigma^{(i)2}]$, where c is a normalisation constant. One can then introduce the distance variables $\delta^{(i)} := (v^{(i)} - \mu^{(i)})/\sigma^{(i)}$, and so the Gaussian becomes $G^{(i)}(v^{(i)}) = c\exp[(\delta^{(i)})^2/2]$. The quantities $\delta^{(i)}$ measure the difference of $v^{(i)}$ from the observational mean in terms of

the observational standard deviation. If, for instance $\delta^{(i)} = 2$ for a specific value of $v^{(i)}$, then this value is two standard deviations away from the observational mean.

Murphy *et al.* then consider a PPE of 53 different model versions of HadSM3, where the model parameters that are varied and their uncertainty ranges are chosen on the basis of expert advice (*ibid*., 769). These model runs produce hindcasts for the $v^{(i)}$, which we denote by '$v^{(i)}(\alpha)$' to indicate that they are produced by a model version defined by specific value of the parameters α. The $G^{(i)}$ can then be used to give a probabilistic weight to the model version in the light of the observations by plugging $\delta^{(i)}(\alpha) = (v^{(i)} - \mu^{(i)}) / \sigma^{(i)}$ into the Gaussians. If for instance, one then finds $\delta^{(31)}(\alpha) = 16$ this means that the value the model version defined by α predicts for variable 31 is 16 standard deviations away from the observational mean. One can then calculate the average of the 32 $\delta^{(i)}(\alpha)$ for a model version. This number is known as the *climate prediction index* (CPI). Murphy *et al.* calculate the CPIs for all 53 model versions and find that it varies between 5 and 8. This means that the average of the model's retrodictions are 5 to 8 standard deviations away from the observations. This, in turn, means that the actual observations are extremely unlikely by the lights of the model! In their Figure 4 (*ibid*., 771) they also give the values for the individual $\delta^{(i)}(\alpha)$. For some variables these values are relatively small (between 1 and 2), but for some variables they range between 23 and 24. Many of the climate variables used by UKCP09 – see Sexton (2012, 2518) for a list – are among the 32 considered in Murphy *et al*. and so this result has implications for UKCP09. If actual observations turn out to be extremely unlikely by the light of the model then why might one hold that the model is quantitatively or probabilistically informative about the climate? It is worth noting that this need not undermine the model's value as a research tool for understanding climatic processes; nor need it imply that the model is uninformative at all temporal and spatial scales for all variables. It does however call into question the validity of the informativeness assumption as a justification for the use of the discrepancy term in the UKCP09 projections.

4.3 The Proxy Assumption

The proxy assumption is the assumption that measuring the average distance of HadSM3 to the 12 members a multi model ensemble yields a result that is close to what one would find if one were to measure HadSM3's distance to the real world. The discussion of this assumption is complicated by the fact the literature on the subject exhibits a certain degree of inconsistency. On the one hand the method is illustrated and advertised as delivering trustworthy results; on the other disclaimers that effectively undermine the crucial assumptions are also included, sometimes parenthetically or deep within technical discussions.[33]

The first reason cited in support of the proxy assumption is that multi model averages give a better representation of climate than any individual model: 'Indeed, the multimodel ensemble mean has been shown to be a more skilful representation of the present-day climate than any individual member (Reichler and Kim 2008)' (Sexton et al. 2012, 2526). While often true in a root-mean-square sense, it is not at all clear what implications this holds for probabilities derived from multimodel ensembles (Smith et al. 2014).

It is acknowledged that 'systematic errors to all current climate models persist' (Sexton et al. 2012, 2526). In making climate projections, the models are being used to extrapolate to a state of the system which has not been seen before. The crucial factor for such a task is whether our tool for producing projections – be it a model or a multi-model mean – can claim to be sufficiently close to reality. In a complex nonlinear system on times longer than a mixing time such closeness must be achieved for all variables which we believe could impact the response of our variables of interest. More fundamentally, two senses of skill appear to be mixed here: (a) the skill of a point forecast like the ensemble mean and (b) the skill of a probabilistic forecast like a Bayesian distribution. In each case, note the contrast between 'more skilful', which is a comparative quality a model possesses relative to other models, and 'skilful', which is an intrinsic quality of a model (referring to its skill in supporting decisions). Being *more skilful* is of little relevance unless the original model truly is skilful. Furthermore, when

[33] An example is (J. Murphy et al. 2010, 63-69).

skilful is defined in terms of a single performance index (Reichler and Kim 2008; James M. Murphy et al. 2004) there are nevertheless significant errors in variables which might be expected to substantially influence changes in the future. No evidence is given that *more skilful* can be equated with *skilful* in terms of a single performance index. The literature on model errors suggests that this is also true for many individual variables of undeniable importance for future climate change.

The second reason mentioned in support of the proxy assumption is that 'the structural errors in different models can be taken to be independent' (J. Murphy et al. 2010, 66) and that therefore the ensemble samples uncertainty well. However, immediately after we are warned that

> 'Whilst there is evidence for a degree of independence […], there is also evidence that some errors are common to all models […], due to shared limitations such as insufficient resolution or the widespread adoption of an imperfect parameterisation scheme. From this perspective, our estimates of discrepancy can be viewed as a likely lower bound to the true level of uncertainty associated with structural model errors.' (J. Murphy et al. 2010, 66)

And then the conclusion is drawn that: 'The main (and inevitable) limitation, however, is that it [the proxy assumption] does not account for the potential impacts of errors common to all climate models used in the prediction' (Sexton et al. 2012, 2516).

If there are common errors the proxy assumption fails, and as all of today's models share the same technological constraints posed by today's computer architecture they inevitably share some common errors such as limitations on the accuracy of topography. Indeed such common errors have been widely acknowledged (see, for instance, Knutti et al. 2010) and studies have demonstrated and discussed the lack of model independence (Jun et al. 2008a, 2008b; Bishop and Abramowitz 2013). Furthermore, the mathematical space of all possible climate models (if there is some such thing) is huge, and there is no reason to believe that the 12 models we de facto work with provide a representative sample, even of the subset of models that could be run under the constraints of today's technology.

For these reasons, the assumption that the use of an MME will accurately quantify the distance to our true target is unjustified. It produces a distribution that is more consistent with the diversity of current models, but which need not reflect the uncertainty in the true future climate or even of our uncertainty in future climate given present day scientific understanding. And *nota bene* that the fear is not so much that the width of the uncertainty distribution is somehow too narrow, but rather that the distribution is simply in the wrong place: the mean of the distribution will shift significantly if the model simulations become realistic. Trying to predict the true climate with structurally wrong models is like trying to predict the trajectory of Mercury with Newtonian models. These models will invariably make misleading (and likely maladaptive) projections beyond some lead time, and these errors cannot be removed by adding a linear discrepancy term derived from other Newtonian models. Tests of internal consistency, or other methods to determine the lead times at which the projections are expected to be misleading would be of significant value.

Echoing Murphy *et al.* (2007) we note that '[i]t is important to stress that our approach to the specification of discrepancy can only be expected to capture a subset of possible structural modelling errors and should be regarded as a lower bound' (James M. Murphy et al. 2007, p. 2011). A lower bound on the discrepancy need neither yield trustworthy projections nor provide a suitable basis for quantitative decision support.

5. Emulation

As we have seen in Section 3, emulation is an intrinsic part of UKCP09. The technique used in UKCP09's scheme is a *Gaussian process emulator* (related to *Kriging*). The leading idea of this approach is to treat $\varphi_t(x;\alpha)$ as an uncertain function and treat the values of $\varphi_t(x;\alpha)$ for any collection of input points $\alpha^{(1)},...,\alpha^{(k)}$ (where k is the number of available model runs) as a multivariate Gaussian. Then a mean function η and a covariance function λ are chosen. The mean function gives a prior expectation of $\varphi_t(x;\alpha)$ for any α and λ gives the prior covariance between the values of $\varphi_t(x;\alpha)$ for different values of α. Typically η and λ contain adjustable parameters and the values

of the results of the model runs are used to adjust these parameters. The result of this process is the distribution $\Psi(x;\alpha)$.

The complicated details of this process need not occupy us here. What matters is that this technique is based on the assumption that the target surface $\varphi_t(x;\alpha)$ exists and is smooth (Sexton et al. 2012, 2523). It also matters that the emulation can be expected to yield correct results only if a sufficiently large sample of points from the target surface is available to train the emulator. We now examine whether these assumptions are satisfied in UKCP09.

There are serious questions regarding the existence of the target surface, and if it exists then its smoothness.[34] The root of this problem lies in choosing the space for the emulation, and the rub lies with initial conditions. The problem comes in the form of a trilemma.

First horn of the trilemma: For mathematical tractability one would like to have a target surface with as few dimensions as possible. This can be achieved by choosing a particular initial condition x_0 and only varying the α. The target surface then has 31 dimensions (it is a mapping from the 31 dimensional parameter space into X). But this surface is uninteresting. HadSM3 is a nonlinear dynamical system and one would expect the trajectory of future climate to depend sensitively on initial conditions. In Section 7.1 we will see that HadSM3 indeed exhibits this kind of sensitivity. Therefore the surface defined by $\varphi_t(x;\alpha)$ and x_0 also varies with x_0. When this variation with x_0 is not considered, any particular surface is incomplete in terms of statistical inference.

Second horn of the trilemma: Staying in the 31-dimensional picture (i.e. the α are the only independent variables), one could try to take the dependence on initial conditions into account by plotting a number of different points – corresponding to different initial conditions – for each value of α. The result of this is a swarm of points in X where a

[34] The challenges to smoothness posed by computations on a digital computer are ignored below. We note in passing that such challenges exist when, for example, a change in the least significant bit of x_0 yield significant changes in the target variable (Lorenz 1968).

number of points are associated with every α. This idea suffers from two problems. The first is that UKCP09 simply does not have these points. Model runs for a number of different initial conditions are not available to sample this swarm informatively. The more fundamental problem is that by associating a swarm of points with every value of α one has left the framework of emulation we started with because there now simply is no surface to emulate! Targeting an aggregate function (like the mean over initial conditions) is unhelpful if one is to maintain the fundamental assumption that one knows the target of the emulator exactly at points where the full model has been run.

Third horn: One can expand the space and consider the Cartesian product of the state space and the parameter space A, and regard $\varphi_t(x;\alpha)$ as a mapping from $X \times A$ into X. This move turns the target surface into an object with tens of thousands of dimensions (simply because X has tens of thousands of dimensions). This renders proper emulation untenable. Attempting to emulate a target surface in a 10,000 odd dimensional space on the basis of around 300 points (not even one point per dimension!) would leave the surface dramatically underdetermined.

The conclusion is that whichever way emulation is done, it does not provide the desired statistical emulation.

And finally there is also a model-world confusion in the use of the emulator even when it is used as a stand in for a single model: the x occurring in the relevant emulations is a model variable not a real-world variable. A perfect emulator could (more) quickly produce a distribution which accurately reflected the probability of the next full model run: not the probability of an event in the world. This is a serious cause for concern as long as the diversity of our model is believed not to reflect the uncertainty in the future of the real world (Smith 2002).

6. The Trapezoid Distribution

When using Equation (5) to determine the $p(\hat{y}_f \mid \tilde{o})$, it is assumed that $p(\alpha)$ is the trapezoid distribution $T(\alpha)$. This choice is crucial: $p(\alpha)$ gives a weight to different model versions, and given that different model versions project different futures shifting these weights could, in principle, produce very different projections. So we have two questions. First, what justifies the choice of a distribution that is flat over the middle 75% of uncertainty intervals and drops linearly at the extreme values in the case of continuous parameters and uniform over all options in the case of discrete parameters? Second, how sensitively do the results depend on these choices?

We have been unable to find a justification for the choice of this distribution in the documentation of UKCP09. However, the reasoning looks familiar: as long as one has no reason to prefer one outcome over the other, one should not arbitrarily favour one outcome over the other and assign equal probabilities to both. This is known as the *Principle of Indifference*, which was given a canonical formulation by Laplace in 1814. The principle can be applied successfully to simple situation such as coin flips (where we have no reason to prefer one side to the other), but it suffers from a number of well-known problems.[35]

In the current context the most significant problem is the fact that the principle can be applied in different ways to the *same* situation, which leads to inconsistent probability assignments. This happens whenever a situation can be characterised by two inter-definable quantities which (a) provide equally good descriptions of the situation and (b) are related to one another by a non-linear definitional relationship. Consider a simple example. Suppose you are a numismatic enthusiast and you will for the first time in your life get to see a rare coin. You have seen pictures of the coin and so you know that it is cylindrical with a height much less than its diameter. But you have no idea about its size. Plausibility considerations – people typically keep coins in their pockets! – lead you to think that its diameter δ must be somewhere between 1cm and 5cm, but you know nothing else about the diameter. So you apply the principle of indifference to the diameter and put an even distribution over the interval $[1,5]$. From this you conclude

[35] The principle and its problems are well documented in the philosophical literature on probability; see, for instance, (Salmon et al. 1992, 74-77).

that the probability of the coin having a diameter between 1cm and 3cm is 0.5. But a coin can equally well be described by the surface area of one side, σ. The diameter and the area are inter-definable: $\sigma = \pi(\delta/2)^2$. So given your assumptions about the minimum and the maximum diameter, you think that the coin will have a surface between (approximately) 0.785cm^2 and 19.625cm^2. Now you apply the principle of indifference to the surface, which yields a uniform distribution over [0.785,19.625]. So your probability that the coin has a surface between [0.785,9.42] is 0.5. So far so good. But a coin with a surface of 9.42cm^2 is also a coin with diameter of 3.46cm. So the principle of indifference tells us both that there is a 0.5 probability for the coin having a diameter between 1cm and 3cm, and that there is a 0.5 probability for the coin having a diameter between 1cm and 3.46cm. Since there is no reason to prefer, say, the diameter to the surface area to describe to coin, no probability assignment is preferred and we are faced with a contradiction.

The same problem comes up in the parameter space of the climate model. At least some parameters (i.e. some components of the vector α) are like diameter in that there are inter-definable quantities which are equally legitimate as a description of the physical situation and which relate to the original parameter by a non-linear function. An example is the so-called *ice fall rate* ρ, which describes how fast ice falls out of model clouds.[36] But the same physical effect could just as well be described by the *ice residence time* τ, measuring how long ice stays within a model cloud. The two quantities are inversely proportional: $\tau \sim 1/\rho$. The situation is now exactly analogous to the coin example. The principle of indifference would suggest that we put an even distribution over certain intervals for both quantities, but these distributions will provide contradictory probabilities.

One might try to mitigate the force of this objection by arguing that the choice of a particular $p(\alpha)$ has no significant effect on the posterior probabilities $p(\bar{y}_f \mid \tilde{o})$. UKCP09 consider this issue (although not as a response to the principle of indifference). They acknowledge that 'alternative and equally defensible prior distributions could be

[36] See (D.A. Stainforth et al. 2005; David A. Stainforth et al. 2007) for a discussion of the ice fall rate.

proposed' but claim that 'the results are quite robust to a number of reasonable alternative choices' (Science Report p. 63). An investigation of this issue is presented in an appendix, where it is pointed out that:

> 'prior distributions are recognised as being themselves uncertain […] so we investigate two other choices: assuming uniform probability across the full expert range, and assuming uniform probabilities across a full range of values 15% larger than that specified by experts. The latter, in particular, is a conservative specification which assumes both that the experts systematically underestimated the extremes of their ranges, and that the extreme values can be assumed no less likely than values near the middle of the range.' (J. Murphy et al. 2010, 143)

It is found that 'the impacts on the posterior projections are more modest, and the induced differences in probability are also relatively small compared with the uncertainties indicated by the UKCP09 distributions' (J. Murphy et al. 2010, 144).

Unfortunately, this is insufficient to set worries about alternative distributions to rest. The stability checks performed only vary the original distributions slightly and do not take radically different distributions into account, for instance the kind of distributions one would get for variables that are inversely related to the parameters used. On the basis of these checks it is impossible to assert that the projections are immune to difficulties arising in connection with the principle of indifference.

7. Further Concerns

Two further aspects of UKCP09's methodology give rise to concerns: the incomplete discussion of initial condition uncertainty and the use of downscaling. We will address these briefly in this section.

7.1 Initial Condition Uncertainty

HadSM3 is a non-linear dynamical system and as such one would expect the trajectory of future climate to depend sensitively on initial conditions. Every model run assumes a particular initial condition and hence varying that initial condition only slightly could have produced very different results. So it is conceivable that varying the initial

conditions of the 297 model runs on which UKCP09 is based just by a little bit would have yielded different simulation results, and as consequence given rise to different probabilistic projections. What reasons are there to believe that this is not the case?

In the introductory parts of the Science Report UKCP09 acknowledges the importance of initial condition uncertainty (J. Murphy et al. 2010, 26-28). They point out that we find 'natural variability' in the climate system: 'Climate, at a global scale and even more at a local scale, can vary substantially from one period (for example, a decade or more) to the next, even in the absence of any human influences.' (*ibid*., 26) This variability is seen as being (at least in part) due to 'the chaotic nature of the climate system' (*ibid*., 26). They submit that effects of natural variability can be explored by running the model for different initial conditions:

> 'By running the climate model many times with different initial conditions (a so-called initial condition ensemble) we can estimate the statistical nature of this natural variability on a range of space and time scales, and hence quantify the consequent uncertainty in projections.' (*ibid*., 26)

UKCP09's initial condition ensemble consists of three members (i.e. three runs of HadCM3 under the same emission scenario but with different initial conditions). The conclusion drawn from an analysis of these three model runs is:

> 'It can be seen that, although each experiment [i.e. model run] shows the same general warming, individual years can be quite different, due to the effect of natural internal variability. If we look at changes at a smaller scale, for example those of winter precipitation over England and Wales […] we see that, although the three projections show similar upward trends of about 20% through the century, they are very different from year to year and even decade to decade.' (*ibid*., 27)

Three figures are produced (*ibid*., 26-7) showing how very different the projections are for the three model runs, driving home the point that initial condition uncertainty is not negligible.

UKCP09 conclude their discussion by observing that '[t]he uncertainty due to projected natural internal variability is included in the overall uncertainty quantified in UKCP09' (*ibid*., 27). The above quotation is the last discussion of initial condition uncertainty in

the science report,[37] however, and no further information is provided elsewhere.[38] So it remains at best unclear whether, and if so how, initial condition uncertainty has been taken into account in the production of UKCP09's probabilistic projections.

One might then conjecture that a distribution built up over time from the trajectory starting in one particular initial condition is equivalent to that resulting from an initial condition ensemble. Daron and Stainforth (2013) gave this assumption a name: the kairodic assumption. The idea behind the kairodic assumption is that while individual trajectories may give rise to very different weather patterns, the overall distribution of weather events which makes up the climate remain the same for all trajectories, and for this reason it is unnecessary to sample initial conditions if one is interested in the climate distribution. There is no reason a priori to expect that a distribution is not itself IC dependent (Lorenz 1968). Arguably, there is every reason to expect it to be the case. The question is: how much does it matter and at what scales? Stainforth *et al.* (2007) show that initial conditions had a large impact on multi-year seasonal averages of large regional averages in HadSM3, and so we might expect them to be important.[39] Deser *et al*. (2012) show that multi-decadal trends over the USA in climate models vary substantially with different initial conditions. Daron and Stainforth (2013) showed that the kairodic assumption is likely to be substantially misleading in a climate change like situation, where distributions are expected to change.

An acknowledgement that averaging does not account for initial condition uncertainty can be found in the Science Report:

> A common way of reducing the effect of uncertainty due to natural variability on the projections is to average changes over a 30-yr period, as we did in the UKCIP02 scenarios (and do again in

[37] Initial conditions are mentioned again on p. 129, but no information beyond what has been said on p. 26-27 is provided.

[38] Initial conditions are briefly mentioned but not discussed in (Sexton et al. 2012) and in the Briefing Report (Jenkins et al. 2009). The Science Report (J. Murphy et al. 2010), a document of over 190 pages, dedicates three pages in the introductory part to the problem of initial condition uncertainty.

[39] The multi-years averages in (2007) were 8 years long compared to 30 years in UKPC09. However, the problem pinpointed by Stainforth et al. is unlikely to disappear by moving from 8 year averages to 30 year averages.

> UKCP09). But even this still allows large differences in patterns of change […]; for example over Birmingham where two of the model experiments project approximately 30% increases, but the other projects just over 10%. The uncertainty due to projected natural internal variability is included in the overall uncertainty quantified in UKCP09.' (*ibid*., 27)

So the accepted wisdom that initial condition uncertainty can be well represented by variability within a single trajectory is untenable, and avoiding a serious exploration of the dependence of climate distributions on initial conditions by appeal to the kairodic assumption is unjustified. The lack of exploration of initial condition uncertainty therefore remains a concern.

7.2 Downscaling

There are many approaches to downscaling, and much could be said about alternative approaches regarding translating the coarse outputs of a GCM to information on the space and time resolution of phenomena of relevance to decision making. Such a discussion is unnecessary here, as it suffices to note that downscaling does not account for significant inadequacies in the coarse resolution data it takes as inputs. This is clearly stated in the internal Hoskins Review of UKCP09:[40]

> 'The focus on UK-scale climate change information should not obscure the fact that the skill of the global climate model is of over-whelming importance. Errors in it, such as the limited current ability to represent European blocking, cannot be compensated by any downscaling or statistical procedures, however complex, and will be reflected in uncertainties on all scales.'

Given the issues we have raised above regarding the interpretation of the GCM, we do not discuss the downscaling step in detail.

There is some confusion regarding the role of GCMs, the computation of global averages and the use of downscaling in UKCP09.[41] In the scheme used in UKCP09 a

[40] A summary of the review is available at http://ukclimateprojections.metoffice.gov.uk/23173. The above quotation has been retrieved on 7 March 2014.

[41] We thank an anonymous referee for raising this concern.

regional climate model (RCM) is coupled to a GCM, and the RCM is used to simulate the local climate using the inputs from the GCM as boundary conditions. The GCMs considered in UKCP09 have coarse regional detail. If the regional detail is badly wrong, then the RCMs driven by that GCM output will produce misleading outputs. This does not necessarily imply, however, that global averages would have to be wrong too if one could somehow calculate such averages using the techniques discussed above. Whether they would is a question that need not occupy us here. It is entirely appropriate to say we are considering high-resolution projections and then focus primarily on the GCMs; focusing on GCMs is in no way synonymous with focusing on global averages.

8. Is UKCP09 Intended For Actual Use?

It has been suggested to us by an anonymous referee that we are attacking a straw man because the authors of the UKCP09 reports are well aware of these limitations and do not intend their results to be used for decision support. The Briefing report acknowledges that the probabilities provided are derived using a number of assumptions and that 'probabilistic estimates are robust to reasonable variations *within these assumptions*' (Jenkins et al. 2009, 6; emphasis added), and it emphasises that 'probabilistic projections are themselves uncertain' (*ibid*., 25).[42] Statements like these, so the referee continues, indicate that UKCP09 is aware of the limitations of their method. Their aim is to provide 'as much good information as possible' (ibid. 9) to decision-makers, and this, so the referee emphasises, is not equivalent to suggesting that UKCP09 provides decision-relevant projections.[43]

UKCP09 places great emphasis on practical applications and on providing evidence for policy makers. We have seen in Section 2 that they aim to offer advice to 'planners and other decision-makers in sectors such as transport, healthcare, water-resources and

[42] Sometimes this observation comes in the guise of there being a cascade of uncertainty, with moderate confidence at the continental scale and less confidence at the local scale; see for instance (Jenkins et al. 2009, 6 and 22). How exactly the point that local projections are uncertain is expressed is immaterial to the dialectic in this section.

[43] These alternative readings are also discussed in Parker (2014).

coastal defences' (Jenkins et al. 2009, 9). The Briefing Report observes that 'The provision of probabilistic projections is the major improvement which the UKCP09 brings to *users*' (*ibid*., 23; emphasis added) and stresses that probabilistic projections 'can help with making robust adaptation decisions' (*ibid*., 23). Finally, the information is used in 'worked examples', where problems like energy use and sustainability in school buildings, overheating risk for buildings, potential changes to snowfall in Snowdonia, UK marine shelf conservation, climate change and forestry adaptation, and flood management policy are assessed using the UKCP09 projections.[44] Plainly, information is offered and treated as decision-relevant.

The observation that projections themselves are uncertain and valid only within the assumptions made is potentially undermining: projections can be decision-relevant only if they are believed to be solid enough to bet on. One cannot at once maintain that projections are decision-relevant, and that they are only model-immanent and lack a definite connection to the real world. UKCP09 implicitly recognise this truism when they reassure the reader that model results are on the right track:

> 'Although it is important that prospective users understand the limitations and caveats, it is also worth emphasising that (a) current models are capable of simulating many aspects of global and regional climate with considerable skill; and (b) they do capture, albeit imperfectly, *all the major physical and biogeochemical processes known to be likely to exert a significant influence on global and regional climate over the next 100 yr or so.*' (Jenkins et al. 2009, 45; emphasis added)

> 'The UKCP09 projections can make a useful contribution to assessing risks posed by future climate*; they are appropriate for informing decisions on adaptation to long-term climate change which need to be taken on the basis of current knowledge* […]' (Jenkins et al. 2009, 46; emphasis added)

Clearly the message is: while there may be inaccuracies, they are small enough not to undermine the practical usefulness of the probability projections. So, contrary to the referee's objection, UKCP09 is committed to the decision-relevance of its projections. We have argued that this commitment is unwarranted.

[44] See http://ukclimateprojections.metoffice.gov.uk/23102 for details.

9. Conclusion

We find little support for interpreting UKCP09's projections as trustworthy projections for quantitative decision support, alongside significant doubts that the information required for such high-resolution projections is at hand today. Needless to say, questioning the evidence for a result does not amount to proving it wrong; our concern is that the premises of the argument do not warrant trust in the results. We suggest that necessary (not sufficient) conditions for a warrant of trust include the provision of:

1. Evidence that the discrepancy term used in practice is sufficiently informative about the real system: specifically that the set of models share no known shortcoming and that the diversity of current models can be taken to capture the uncertainty in the true future climate.
2. Evidence that the form of the discrepancy term (in the case of UKCP09, Gaussian) is sufficient to capture the structural errors that are sampled.
3. Evidence that the Bayesian priors adopted (in the case of UKCP09 the trapezoid distribution) yield a robust outcome (under reasonable changes to the prior) and a relevant posterior (relevant to the risk management targeted).
4. Evidence that the emulator provides the desired statistical emulation; in particular that the emulation problem is well-posed given the nature of the system as reflected in the model being emulated (nonlinear, chaotic, and so on), and that the emulation is effective out-of-sample.
5. Evidence that every known uncertainty has been accounted for (in the case of UKCP09 initial condition uncertainty is neither sampled nor is its impact reflected in the projections).
6. Evidence that each translation from model variables to corresponding variables in the world (that is, from a future model state to a decision relevant quantity we will observe in the future[45] accounts for all known model shortcomings, that those lost in translation are captured in the discrepancy, and that the implications of those missed are made clear. (While both the limited adequacy of

[45] Discussion of the translation between model variables and real world variables with similar names can be found in Smith (2000) and Smith (2002).

downscaling and the low fidelity in simulated blocking are openly acknowledged by UKCP09 in this regard, the implications these inadequacies hold for decision-making are not made clear.)

7. Evidence and argument supporting a minimum lead time on which the process has a strong warrant of trust, and a maximum lead time beyond which projections are not expected to be trustworthy. These time-scales will, of course, vary with space and duration of the target variables.

These seven elements stand individually. While resolving any one of them in the context of climate would be a major research achievement, the shift from being a valuable research programme which advances science to a trustworthy operational risk management tool for decision makers requires resolving each one. We do not claim this list is complete. In the context of UKCP09, each element on its own is sufficiently worrisome to cast doubt on the decision-relevance of the information as quantitative risk-management tool. On the available evidence UKCP09's projections do not merit trust.

In the case of decision support in the face of climate change this is a crucial point for two reasons. First, over-reliance on the reliability of such climate projections can undermine the ability to make robust decisions; better decisions could be made with a better understanding of the scientific uncertainties even when they cannot be presented in this quantitative fashion. Second, for the reasons outlined above, the detailed probabilistic projections might be expected to change [46] substantially in future assessments, thus undermining the user communities' trust in scientific outputs; particularly their presentation of uncertainty.

It should be noted that the scientists who worked hard to make UKCP09 the best it could be were constrained by the structure of the project; the deliverables were defined before any viable approach to meet them was available in the peer-reviewed literature. Furthermore, the *United Kingdom Climate Impacts Program*, which is much broader

[46] Expected to change even without a deeper scientific understanding of the phenomena, or new observations. Scientific projects are always subject to change when our basic understanding of science changes, the question is whether they are mature conditioned on everything we know today.

than UKCP09, faced the problem of motivating users to engage with the real challenges and risks posed by climate change in the face of deep uncertainty regarding local impacts: the challenge of keeping users engaged and interested when the information they most immediately desire may lie beyond the reach of today's science.

Pointers to the fact that a naïve interpretation of UKCP09 probability distributions is untenable can be found within the UKCP09 material. The UKCP09 worked examples, however, clearly suggest the decision-making application of this material in ways which, if our criticisms hold true, would be expected to prove maladaptive.

When the best available tool in terms of the utility of its deliverables is not adequate for purpose (trustworthy), it is not in fact 'best available'. In this case good policy, decision making, and risk management would be based on trustworthy, if less desirable, deliverables. Where tools like UKCP09 are not trustworthy, what is? The aim of this paper was to pave the ground for an informed discussion of this question. As long as the prevailing view is that (something like) the probabilities provided by UKCP09 offer an attainable trustworthy option today, the issue of more informative approaches does not even arise. We hope we have illuminated a way to move forward to new horizons.

Acknowledgements

Work for this paper has been supported by the LSE's *Grantham Research Institute on Climate Change and the Environment* and the *Centre for Climate Change Economics and Policy* funded by the Economics and Social Science Research Council and Munich Re. Frigg further acknowledges financial support from the AHRC-funded 'Managing Severe Uncertainty' project and grant FFI2012-37354 of the Spanish Ministry of Science and Innovation (MICINN). Smith would also like to acknowledge continuing support from Pembroke College, Oxford. We would like to thank Wendy Parker, Erica Thompson, and Charlotte Werndl for comments on earlier drafts and/or helpful discussions.

References

Allen, M. R., & Stainforth, D. A. (2002). Towards objective probabalistic climate forecasting. *Nature, 419*(6903), 228-228.

Beven, K. (2012). Causal models as multiple working hypotheses about environmental processes. *Comptes Rendus Geoscience*, *344*, 77-88.

Bishop, C. H., & Abramowitz, G. (2013). Climate model ependence and the replicate earth paradigm. *Climate Dynamics, 41*, 885-900.

Daron, J. D., & Stainforth, D. A. (2013). On predicting climate under climate change. *Environmental Research Letters, 8*, 1-8.

Deser, C., Knutti, R., Solomon, S., & Phillips, A. S. (2012). Communication of the role of natural variability in future north American climate. *Nature Climate Change, 2*(November), 775-779.

Dessler, A. (2011). *Introduction to modern climate change*. Cambridge: Cambridge University Press

Du, H., & Smith, L. A. (2012). Parameter estimation through ignorance. *Physical Review E, 86*(1), 016213.

Frigg, R., Bradley, S., Du, H., & Smith, L. A. (2014). The adventures of Laplace's demon and his apprentices. *Philosophy of Science, 81*(1), 31–59.

Jenkins, G., Murphy, J., Sexton, D., Lowe, J., & Jones, P. (2009). UK Climate Projections briefing report. DEFRA.

Judd, K., & Smith, L. A. (2004). Indistinguishable states II: The imperfect model scenario. *Physica D, 196*, 224-242.

Jun, M. Y., Knutti, R., & Nychka, D. W. (2008a). Spatial analysis to quantify numerical model bias and dependence: How many climate models are there? *Journal of the American Statistical Association, 103*, 934-947.

Jun, M. Y., Knutti, R., & Nychka, D. W. (2008b). Local eigenvalue analysis of CMIP3 climate model errors. *Tellus A: Dynamic Meteorology and Oceanography, 60*, 992-1000.

Kennedy, M. C., & O'Hagan, A. (2001). Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63*(3), 425–464.

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

Lorenz, E. (1968). Climate determinism. *Meteorological Monographs, 8*(30), 1-3.

McWilliams, J. C. (2007). Irreducible imprecision in atmospheric and oceanic simulations. *Proceedings of the National Aacdemy of Sciences, 104*(21), 8709–8713.

Meehl, G. A., Goddard, L., Murphy, J., Stoufer, R. J., Boer, G., Danabasoglu, G., et al. (2009). Decadal prediction can it be skillful? *Bulletin of the American Meteorological Society, 90*, 1467ff.

Murphy, J., Sexton, D., Jenkins, G., Boorman, P., Booth, B., Brown, K., et al. (2010). UK Climate Projections science report: climate change projections. Version 3, updated December 2010. http://ukclimateprojections.defra.gov.uk/22544 Met Office Hadley Centre.

Murphy, J. M., Booth, B. B. B., Collins, M., Harris, G. R., Sexton, D. M. H., & Webb, M. J. (2007). A methodology for probabilistic predictions of regional climate change for perturbed physics ensembles. *Philosophical Transactions of the Royal Society A, 365*, 1993-2028.

Murphy, J. M., Sexton, D. M. H., Barnett, D. N., Jones, G. S., Webb, M. J., Collins, M., et al. (2004). Quantification of modelling uncertainties in a large ensemble of climate change simulations. *Nature, 430*(12 Agust), 768-772.

Oreskes, N. (2007). The scientific consensus on climate change: How do we know we're not wrong? In J. F. C. DiMento, & P. Doughman (Eds.), *Climate change: What it means for us, our children, and our grandchildren* (pp. 65-99). Boston: MIT Press.

Oreskes, N., Stainforth, D. A., & Smith, L. A. (2010). Adaptation to global warming: Do climate models tell us what we need to know? *Philosophy of Science, 77*(5), 1012-1028.

Parker, W. (2010a). Predicting weather and climate: Uncertainty, ensembles and probability. *Studies in History and Philosophy of Modern Physics, 41*(3), 263-272.

Parker, W. (2010b). Whose probabilities? Predicting climate change with ensembles of models. *Philosophy of Science, 77*(5), 985-997.

Parker, W. (2013). Ensemble modeling, uncertainty and robust predictions. *Wiley Interdisciplinary Reviews: Climate Change, 4*(3), 213-223.

Parker, W. (2014). Values and uncertainties in climate prediction, revisited. *Studies in History and Philosophy of Science*, *46*, 24-30.

Reichler, T., & Kim, J. (2008). How well do coupled models simulate today's climate? *Bulletin of the American Meteorological Society, 89*(3), 303-311.

Rougier, J. (2007). Probabilistic inference for future climte using an ensemble of climate model evaluations. *Climatic Change, 81*, 247-264.

Rougier, J. (2008). Efficient emulators for multivariate deterministic functions. *Journal of Computational and Graphical Statistics, 17*(4), 27-843.

Salmon, M., Earman, J., Glymour, C., Lennox, J. G., Machamer, P., McGuire, J. E., et al. (1992). *Introduction to the philosophy of science*. Indianapolis and Cambridge: Hackett.

Seager, R., Kushnir, Y., Ting, M. F., Cane, M., Naik, N., & Miller, J. (2008). Would advance knowledge of 1930s SSTs have allowed prediction of the Dust Bowl drought? *Journal of Climate, 21*, 3261-3281.

Sexton, D. M. H., & Murphy, J. M. (2012). Multivariate probabilistic projections using imperfect climate models part II: Robustness of methodological choices and consequences for climate sensitivity. *Climate Dynamics, 38*, 2543-2558.

Sexton, D. M. H., Murphy, J. M., Collins, M., & Webb, M. J. (2012). Multivariate probabilistic projections using imperfect climate models part I: Outline of methodology. *Climate Dynamics, 38*, 2513-2542.

Smith, L. A. (2000). Disentangling uncertainty and error: on the predictability of nonlinear systems. In Mees, A.I. (ed.) Nonlinear Dynamics and Statistics, Boston: Birkhauser, 31-64.

Smith, L. A. (2002). What might we learn from climate forecasts? *Proceedings of the National Academy of Science, USA 4*(99), 2487-2492.

Smith, L. A. (2006). Predictability past predictability present. In T. Palmer, & R. Hagedorn (Eds.), *Predictability of Weather and Climate* (pp. 217-250). Cambridge: Cambridge University Press.

Smith, L. A., Du, H., Suckling, E. B., & Niehörster, F. (2014). Probabilistic skill in ensemble seasonal forecasts. *Quarterly Journal of the Royal Meteorological Society, 140*(684), DOI:10.1002/qj.2403.

Smith, L. A., & Stern, N. (2011). Uncertainty in science and its role in climate policy. *Philosophical Transactions of the Royal Society A, 369*, 1-24

Solomon, S., Qin, D., & Manning, M. (Eds.). (2007). *Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge: Cambridge University Press.

Stainforth, D. A., Aina, T., Christensen, C., Collins, M., Faull, N., Frame, D. J., et al. (2005). Uncertainty in predictions of the climate response to rising levels of greenhouse gases. *Nature, 433*(7024), 403-406.

Stainforth, D. A., Allen, M. R., Tredger, E. R., & Smith, L. A. (2007). Confidence, uncertainty and decision-dupport relevance in climate predictions. *Philosophical Transaction of the Royal Socity A, 365*(1857), 2145-2161.

Stocker, T. F., Qin, D., Plattner, G.-K., Tignor, M. M. B., Allen , S. K., Boschung, J., et al. (Eds.). (2013). *Climate change 2013. The physical science basis. Working Group I contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge: Cambridge University Press.

Tang, S., & Dessai, S. (2012). Usable science? The UK Climate Projections 2009 and decision support for adaptation planning. *forthcoming in Weather, Climate, and Society*.

Thompson, E. L. (2013). *Modelling North Atlantic Storms in a Changing Climate. Ph.D. Thesis*: Imperial College, London, UK.

Winsberg, E. (2012). Values and uncertainties in the predictions of global climate models. *Kennedy Institute of Ethics Journal, 22*(2), 111-137.

Winsberg, E., & Biddle, J. (2010). Value judgements and the estimation of uncertainty in climate modeling. In P. D. Magnus, & J. B. Busch (Eds.), *New waves in philosophy of science* (pp. 172-197). London: Palgrave Macmillan.