Brit. J. Phil. Sci. **64** (2013), 609–635

Climate Models, Calibration, and Confirmation

Katie Steele and Charlotte Werndl[†]

ABSTRACT

We argue that concerns about double-counting—using the same evidence both to calibrate or tune climate models and also to confirm or verify that the models are adequate—deserve more careful scrutiny in climate modelling circles. It is widely held that double-counting is bad and that separate data must be used for calibration and confirmation. We show that this is far from obviously true, and that climate scientists may be confusing their targets. Our analysis turns on a Bayesian/relative-likelihood approach to incremental confirmation. According to this approach, double-counting is entirely proper. We go on to discuss plausible difficulties with calibrating climate models, and we distinguish more and less ambitious notions of confirmation. Strong claims of confirmation may not, in many cases, be warranted, but it would be a mistake to regard double-counting as the culprit.

[†] Authors are listed alphabetically; this work is fully collaborative.

doi:10.1093/bjps/axs036

Advance Access published on February 28, 2013

1 Introduction

Climate scientists express concern about the practice of 'calibrating' climate models to observational data (another widely used word for 'calibration' is 'tuning'). Calibration occurs when a model includes parameters or forcings about which there is much uncertainty, and the value of the parameter or forcing is determined by finding best fit with the data. That is, the parameter or forcing in question is effectively a free parameter, and calibration determines which value(s) for the free parameter best explain(s) the data. A prominent example, which we refer to later, is the fitting of the aerosol forcing.

The apparent concern about calibration is that it may result or always results in data being double-counted: data used to construct the fully-specified model are also used to evaluate the model's accuracy, in a problematic way. Indeed, various climate scientists worry about circular reasoning:

> In addition some commentators feel that there is an unscientific circularity in some of the arguments provided by GCMers [general circulation modelers]; for example, the claim that GCMs may produce a good simulation sits uneasily with the fact that important aspects of the simulation rely upon [...] tuning. (Shackley *et al.* [1998], p. 170)

This is one particularly suggestive quote about the badness of double-counting. But what exactly is the badness here? We will see that this depends crucially on the details.

This article seeks to clarify and evaluate worries surrounding calibration and double-counting. We appeal to statements made by various climate scientists, but our aim is not to rebut particular individuals. Our main concern is that, in general, climate scientists' statements about calibration/tuning/double-counting do not attend to the details, and are, at worst, misleading. A number of different issues are bundled together as the 'problem of double-counting', and each of these issues deserves to be carefully articulated.

It is necessary to introduce some terminology. Calibration is introduced above. Confirmation refers to the evaluation of a model's accuracy for particular purposes.[1] Note also that there is an important difference between incremental and absolute confirmation: the former concerns whether confidence in a model hypothesis has increased, whereas the latter concerns whether confidence in a model hypothesis is sufficient, or above some threshold. This article focuses on (varieties of) incremental confirmation.[2] The

[1] Some authors, for example, Frame *et al.* ([2007]), use the term 'verification' in lieu of 'confirmation'. We use the latter term in the interests of making a clear connection with the philosophical literature.

[2] From now on, when we use the term 'confirmation' we mean incremental confirmation, unless otherwise indicated. As will become clear, we distinguish two varieties of incremental confirmation: comparative and non-comparative.

central question is: what is the inevitable/proper relationship between calibration and confirmation?

Some climate scientists appear to claim that calibration is bad and should therefore be avoided:

> Climate change simulations should, in general, only incorporate forcings for which the magnitude and uncertainty have been quantified using basic physical and chemical principles. (Rodhe *et al.* [2000], p. 421)

This statement may be stronger than the authors intend.[3] In any case, an anti-calibration position is not defensible, because it would preclude refining models in response to observational evidence. This is common practice in all areas of science. In short, whatever the details of the relationship between calibration and confirmation, it had better be the case that calibration is not something that is bad or needs to be avoided.

Our main target here is the widespread view amongst climate scientists that calibration and confirmation should be kept 'separate'. The following quotes suggest that evidence used in calibration should not (or cannot) yield incremental confirmation; only separate data, not already used for calibration, can boost confidence in a model. In other words, tuning is fine if it simply amounts to calibration, but double-counting is not fine:

> The inverse calculations [calibration] are also based on sound physical principles. However, to the extent that climate models rely on inverse calculations, the possibility of circular reasoning arises—that is, using the temperature record to derive a key input to climate models that are then tested against the temperature record. (Anderson *et al.* [2003], p. 1103).

> If the model has been tuned to give a good representation of a particular observed quantity, the agreement with that observation cannot be used to build confidence in that model. (Randall and Wood [2007], p. 596)

Indeed, the need for separate data for calibration and confirmation is usually simply taken for granted in the climate science literature, or else the reasoning is ambiguous.[4] But this position is far from being obviously true, and requires further argument.

The first part of the article argues that separate data for calibration and confirmation are not an uncontroversial tenet of confirmation logic, because it does not follow (in fact, quite the contrary) from at least one major approach

[3] Perhaps the authors want to exclude forcings that have no physical plausibility at all, rather than forcings that merely cannot be well quantified.

[4] See, for instance, Anderson *et al.* ([2003], p. 1103), Knutti ([2008], p. 4651; [2010], p. 399), Randall and Wood ([2007], p. 596), Shackley *et al.* ([1998], p. 170) and Tebaldi and Knutti ([2007], p. 2070).

to confirmation—the Bayesian approach.[5] After some remarks in Section 2 about climate models and adequacy-for-purpose that are useful to bear in mind throughout the discussion, in Sections 3 and 4 we demonstrate, using a very basic model and examples from climate science, that evidence may be used to calibrate and also to incrementally confirm a model relative to another model (we call this comparative confirmation).

We then go on to address some complicating issues—reasons why in some contexts data are useless for calibration or confirmation. Some climate scientists' worries about double-counting are most charitably reconstructed along these lines, i.e. as concerning the inapplicability, rather than the inherent badness of double-counting. Section 5 considers the issue of 'old evidence'—if evidence already informs the prior probability distribution over models, it cannot be used a second time over for further calibration and confirmation. Section 6 discusses the worry that past data are irrelevant for model adequacy in the future and, hence, cannot be used for calibration or confirmation.

Section 7 discusses a different sense of incremental confirmation that climate scientists may have in mind: non-comparative confirmation, which concerns our confidence in a model *tout court*, i.e. relative to its entire complement. While evidence may also be used to calibrate and confirm a model for non-comparative confirmation, the worry arises that climate models are based on assumptions that may be wrong, especially in the future. Hence, there is considerable uncertainty about the full space of models, implying that data will not confirm a model. Section 8 presents an example from climate science that brings these subtler issues to the fore. The article ends with a conclusion in Section 9.

Let us now turn to the remarks about the predictive purposes of climate models and how this bears on what evidence is relevant for assessing them.

2 Remarks about Models and Adequacy-for-Purpose

A variety of climate models are used to study the Earth's climate. In the words of Parker ([2010], p. 1084):

> [Climate models] range from the highly simplified to the extremely complex and are constructed with the goal of simulating in greater or lesser detail the transport of mass, energy, moisture, and other quantities by various processes in the climate system. These processes include the movement of large-scale weather systems, the formation of clouds and

[5] We try to deal minimally in Bayesian assumptions that may be objectionable to some readers, chiefly, prior probabilities. While we restrict our attention to Bayesian confirmation logic, the lessons apply more broadly, and we note this where appropriate. In any case, our aim is simply to show that it is not uncontroversial to claim that separate data must be used for calibration and confirmation.

> precipitation, ocean currents, the melting of sea ice, the absorption and emission of radiation by atmospheric gases, and many others.

Climate scientists note that many of the aforementioned processes are still poorly understood, and, moreover, that these processes can typically be only approximated in a model, even one of maximum possible precision. Consequently, it is clear from the outset that climate models will not correctly represent or predict the target systems in all their details. This means that climate models themselves cannot be confirmed. As Parker ([2009]) has convincingly argued, instead what can be confirmed is the adequacy of climate models for particular purposes. The hypotheses about the purposes of climate models need to be specified by climate scientists. Here is a prime example of such a hypothesis: this climate model with these initial conditions is adequate for predicting the mean surface temperature changes within 0.5 degrees in the next fifty years under this emission scenario.

In climate science typically some model error is allowed. Therefore, an important part of specifying the hypothesis about the purpose of a model is to state the assumptions about the model error. There are two main kinds of error. First, for discrete model error all that counts is whether the actual outcome is within a certain distance from the simulated outcome, for example, whether the actual and simulated mean surface temperature is $<0.5^{\circ}$C apart. Second, there is probabilistic model error when the error is described by a probability distribution. To give a simple example, the error might be modelled by a Gaussian distribution around the true value.[6]

In this framework of adequacy-for-purpose, one needs to be cautious about what data are actually relevant to assess whether a model fulfills a particular purpose. We have to determine the observational consequences that are likely to follow if the model is adequate; the data about these consequences will then be relevant. To come back to our example about mean surface temperature changes, here many will regard past temperature changes as relevant (although we return to this issue later in Section 6). However, it is less clear whether, for example, past precipitation changes are relevant. As Parker ([2009]) has argued, if climate scientists have obtained a good understanding of the relation between mean surface temperature changes and precipitation changes, then precipitation changes will be relevant. However, when lacking any knowledge about the interdependence of these two variables, then precipitation changes will not be relevant. Which data are relevant is crucial for two reasons: only relevant data can confirm or disconfirm the adequacy of a

[6] One source of error derives from the uncertainty about the initial conditions (this is the problem of internal variability—see Randall and Wood [2007]). It goes without saying that the comparison of models with data also depends on the assumptions made about internal variability. However, this is different from, and does not seem to shed any further light on, the issue of calibration/double-counting. Hence, it will not be discussed further here.

model and can meaningfully be used to calibrate the free parameters of a model.

This article does not, for the most part, focus on the question of what data are relevant to assess a model's adequacy for purpose. General points about the suitability of data for confirmation will, however, become important in Sections 5 and 6. Here, it is just important to realize that this question is a separate issue and should not be confused with the worry of double-counting. That is, if data are not relevant to a model's adequacy for purpose, then testing the model against the data even once would be counting the data one too many times; likewise, calibrating the free parameters of the model against the data would be counting the data one too many times.

The next section discusses calibration/double-counting in the context of more simple models. The aim is to elucidate calibration vis-à-vis Bayesian confirmation.

3 Evidence for Calibration Can Also Yield Comparative Confirmation

Here, we argue against the view that double-counting, in the sense of using evidence for both calibration and confirmation, is obviously bad practice. We show that, by Bayesian or likelihoodist standards at least, double-counting simply amounts to using evidence in a regular and proper way. This is best demonstrated in the context of comparing two well-specified hypotheses. We distinguish two interpretations of double-counting—I (Section 3.1) and II (Section 3.2)—because the legitimacy of the latter is more controversial than the former.

3.1 Double-counting I

Let us start with a straightforward case, and then add complexity. Consider just one type of base model with very simple structure: a linear relationship between variables y and t. Because, as outlined in the previous section, climate scientists typically allow for model error, we will assume a probabilistic model error term that is distributed normally with standard deviation σ:[7]

$$L : y(t) = \alpha t + \beta + N(0, \sigma). \tag{1}$$

The Bayesian account of model calibration depends crucially on the following setup: there is a whole family of specific instances of the base model L,

[7] Alternatively, the error term could be interpreted as observational error or as a combined term for observational error and model error. We focus on model error because it seems particularly widespread in climate science papers. However, all we say carries over to any other interpretation of the error term.

where each specific instance has particular values for the unknown parameters or forcings, α and β. For instance, assume that possible values for α are $\{1, 2, 3, 4\}$, and likewise for β. So the scientist associates with L a (discrete) set of specific model instances that we might label $L_{1,1}$, $L_{1,2}, \ldots$, where the subscripts indicate the values for α and β.

Calibration of L then amounts to comparing specific instances of the base model—$L_{1,1}$, $L_{1,2}$, …—with respect to the data, i.e. observed values for $y(t)$. Of course, strictly speaking, what we are comparing are model hypotheses; assume that the hypotheses here postulate that the model in question accurately describes the data generation process for $y(t)$. Calibration is simply the common practice of testing hypotheses against evidence. Given the probabilistic error term, none of the hypotheses, $L_{1,1}, L_{1,2}, \ldots$, can be falsified by the data, even if the data lies very far away from the specified line. Note also that since the model error is probabilistic, the hypotheses are mutually exclusive. This is important: calibration is best understood as the comparison, given new evidence, of the mutually exclusive hypotheses constituting a base model.[8]

Calibration, understood in this way, may well result in confirmation of $L_{i,j}$, say, with respect to $L_{k,l}$. By Bayesian logic, the extent of confirmation depends on the likelihood ratio: $Pr(E|L_{i,j})/Pr(E|L_{k,l})$, where $Pr(E|L_{i,j})$ is just the probability, Pr, of the evidence, E, i.e. the observed data points, given the model $L_{i,j}$.[9] The likelihoods are related, in a manner that depends on the assumed error probability distribution (in our case Gaussian), to the sum-of-squares distance of the data points from the line. If the likelihood ratio is >1, then $L_{i,j}$ is confirmed by the data relative to $L_{k,l}$, and vice versa if the likelihood ratio is <1. When the likelihood ratio equals 1, neither hypothesis is confirmed relative to the other. Note that the relative posterior (post-evidence) probabilities of $L_{i,j}$ and $L_{k,l}$ are a further matter of absolute rather than incremental confirmation (cf. comments in Section 1); absolute confirmation depends also on their relative prior (initial) probabilities.[10]

[8] Where model error is discrete, identifying mutually exclusive model hypotheses is more complicated. For instance, consider a simple example of two hypotheses involving discrete model error: $L_{1,1}$ is the hypothesis that $y(t) = t + 1$ accurately predicts $y(t)$ within ± 2, and $L_{1,2}$ is the hypothesis that $y(t) = t + 2$ accurately predicts $y(t)$ within ± 2. These two hypotheses could both be correct. Indeed, the model hypotheses in Knutti *et al.* ([2002], [2003]) discussed later in Section 4 and 7 deserve further scrutiny on this basis. We will not discuss this further here; we merely want to flag the issue.

[9] To be more precise, we should also explicitly state the background knowledge, B, in the likelihood expressions, such that they read $Pr(E|L_{i,j}$ and $B)$. In the interests of readability, we will not use these more precise expressions, but the B should be understood as implicit.

[10] This is the Bayesian wisdom, anyhow. The complete Bayesian expression is as follows:

$$\frac{Pr_f(L_{i,j})}{Pr_f(L_{k,l})} = \frac{Pr(L_{i,j}|E)}{Pr(L_{k,l}|E)} = \frac{Pr(E|L_{i,j})}{Pr(E|L_{k,l})} \times \frac{Pr(L_{i,j})}{Pr(L_{k,l})} \quad (2)$$

where the first term is the ratio of posterior probabilities, i.e. the ratio of probabilities after receipt of the evidence. The final term is the ratio of prior or initial probabilities for the model hypotheses, i.e. before the evidence. In short, the ratio of posteriors for the model hypotheses,

We begin with this case to show that there is a straightforward way in which double-counting is fine. Calibration of L involves ascertaining appropriate values for α and β; thus, the whole point is to consider which specific model hypotheses are confirmed relative to others in light of the data. Call this double-counting I; we do not expect its legitimacy to be controversial, given a hypothesis space as described above. So we already see that unqualified statements about the badness of calibration/double-counting are problematic. Note that, for double-counting I, calibration can be regarded as the same process as confirmation in the sense that the evidence is used to do both calibration and confirmation simultaneously: different model instances are considered, and then the evidence is used once to confirm model instances over others. Hence the term 'double-counting' here does not signify that the evidence is used twice; just that it is used for both calibration and confirmation.

3.2 Double-counting II

An interesting qualification may be deduced from the work of Worrall ([2010]). He suggests that the real double-counting sin would be to use evidence to calibrate a base model such as L above, and also hold that the same evidence confirms not only specific instances of this base model relative to others but also the base-model hypothesis itself:

> Using empirical data e to construct a specific theory T' within an already accepted general framework T leads to a T' that is indeed (generally maximally) supported by e; but e will not, in such case, supply any support at all for the underlying general theory T. (Worrall [2010], p. 143)

Call this double-counting II. In this quote, Worrall refers to a general theory, T, that is already 'accepted'. In such a case, the general theory cannot be incrementally confirmed, as it already has maximal probability.[11] Worrall's remarks are thus consistent with Bayesian confirmation. We take Worrall's work to be highly suggestive, however, of the more general claim against double-counting II. We will show that, according to Bayesian confirmation theory, double-counting II is legitimate—thus conflicting with the more general claim against double-counting II.

Perhaps when climate scientists claim that separate data are required for confirmation and calibration, they take for granted, along the lines of Worrall,

given new evidence, E, is a product of the ratio of prior probabilities and the likelihood ratio. As mentioned, it is the likelihood ratio that governs the relative extent to which the model hypotheses are confirmed by E. Note that the likelihood ratio plays a key role in other theories of confirmation too, not just the Bayesian.

[11] Note also that Worrall considers only cases where the evidence falsifies all but one instance of a base model.

that double-counting II is illegitimate, i.e. calibration of a base-model hypothesis cannot result in that hypothesis being confirmed relative to another base-model hypothesis, and thus other data are needed for any such confirmation.

This position, however, is not born out by Bayesian confirmation logic (at least).[12] On the contrary, double-counting II is legitimate and can arise for two reasons: (i) 'average' fit with the evidence may be better for one base model relative to another, and/or (ii) the specific instances of one base model that are favoured by the evidence may be more plausible than those of the other base model that are favoured by the evidence.[13]

As per double-counting I, our analysis revolves around straightforward likelihood ratios, although here we must introduce prior probability distributions over the specific model instances, conditional on each base-model hypothesis being true.[14] In the interests of a more concrete discussion, we first introduce a second base-model hypothesis, a quadratic of the form:

$$Q : y(t) = \alpha t^2 + \beta + N(0, \sigma). \tag{3}$$

Assume that the specific model instances, like those of L above, are all combinations of α and β, where each may take any value in the discrete set $\{1, 2, 3, 4\}$. As before, the error standard deviation, σ, is fixed. Specific model instances are labelled $Q_{1,1}$, $Q_{1,2}$, ... Note that the base-model hypotheses, L and Q, are of the same complexity, i.e. they have the same number of free parameters. This is an intentional choice. We do not want to introduce a further issue of relative model complexity and penalties for overfitting. While an important and controversial issue that is certainly tied up with calibration, the overfitting debate only confounds the question of double-counting. (Nonetheless, we will return to this debate briefly at the end of the subsection.)

In standard Bayesian terms, the confirmation of one base-model hypothesis, for example, L, with respect to another, for example, Q, depends on the likelihood ratio $Pr(E|L)/Pr(E|Q)$. As before, if the ratio is >1, then L is

[12] We remark on frequentist 'model selection' methods at the end of this section. According to these methods, double-counting II is legitimate, in conflict with the general claim we are attributing to Worrall. Note that Mayo's 'severe testing' approach to confirmation does not support the Worrall conclusion either (see Mayo's ([2010]) response to Worrall). What is important for the severe testing approach is not whether evidence has already been used to calibrate a base-model, but whether the evidence severely tests this base-model hypothesis. These two considerations do not always match up. It is beyond the scope and aims of this paper, however, to elaborate further on the severe testing approach or any other alternative vis-à-vis Bayesian confirmation.

[13] Our analysis is thus more in line with Howson ([1988]).

[14] For double-counting I, we were able to eschew prior probabilities altogether when assessing confirmation.

confirmed relative to Q, and if it is <1, then Q is confirmed relative to L.[15] In this case, the relevant likelihoods, however, are not entirely straightforward:

$$\begin{aligned} Pr(E|L) &= Pr(E|L_{1,1}) \times Pr(L_{1,1}|L) + \cdots + Pr(E|L_{4,4}) \times Pr(L_{4,4}|L), \\ Pr(E|Q) &= Pr(E|Q_{1,1}) \times Pr(Q_{1,1}|Q) + \cdots + Pr(E|Q_{4,4}) \times Pr(Q_{4,4}|Q). \end{aligned} \quad (4)$$

Note that $Pr(L_{1,1}|L)$ is the prior probability (i.e. probability before the data is received) of $y(t) = t + 1 + N(0, \sigma)$ being the true description of the data generation process for $y(t)$, given that the true model is linear. The expressions above provide formal support for our earlier statement that confirmation of base models depends on (1) fit with the evidence and (2) the conditional priors of all specific instances of these base models.

Consider first the special case where the conditional prior probabilities of all specific instances of L and Q are equivalent. That is:

$$Pr(L_{1,1}|L) = \cdots = Pr(L_{4,4}|L) = \cdots = Pr(Q_{1,1}|Q) = \cdots = Pr(Q_{4,4}|Q) = x. \quad (5)$$

Suppose the observed data, E, yield on balance greater likelihoods for instances of L than Q. Then L is confirmed relative to Q because of reason (1), viz. the average fit with the evidence is better for base-model hypothesis L than for Q. Furthermore, there is calibration because E is used to determine the most likely values of α and β.

Another special case is where the base-model hypotheses have equivalent fit with the data when all specific models are weighted equally, but the priors are not in fact equal. Suppose that the specific instances of L that have the higher likelihoods for E are in fact more plausible (higher conditional priors) than the specific instances of Q that have the higher likelihoods. Then L is confirmed relative to Q because of reason (2), viz. the specific instances of L favoured by the evidence are more plausible than the specific instances of Q favoured by evidence. Furthermore, there is calibration: E is used to determine the most likely values of α and β.

Alongside these two special cases there is also the case of double-counting II because of both (1) and (2). Note that for double-counting II, as per double-counting I, calibration can be regarded as the same process as confirmation in the sense that the evidence is used to do both calibration and confirmation simultaneously. The evidence is used once to confirm model instances over others, which can then result in base models being confirmed over others.

Worrall ([2010]) has claimed that, in cases where data seem to be used for calibration and confirmation of a base-model hypothesis, what really happens

[15] Again, as before, the relative posterior probabilities of L and Q, i.e. $Pr(L|E)/Pr(Q|E)$, depend also on their prior probability ratio.

is that only some of the data are needed to determine the values of the initial free parameters, and the rest of the data then confirms the hypothesis; thus, there is no double-counting. However, this splitting of the data can throw away valuable information about the free parameters and is not in keeping with Bayesian logic of confirmation. Rather, as we see for the cases discussed here, all of the data are used to determine the values of the free parameters and for confirmation of base-model hypotheses, and thus we have a genuine case of double-counting.

Finally, while the Bayesian approach to confirmation is far from marginal, there have been interesting challenges to this approach in the context of double-counting II. Concerns about comparing base models of differing complexity have lead to special methods for assessing base models, i.e. families of models. This is the field of model selection (see Burnham and Anderson [1998]). Our analysis above is standard Bayesian, but it is important to note that various alternative methods for comparing base models have been suggested, including the Akaike approach (see Forster and Sober [1994]). The controversies here run deep and extend to whether the basic unit of analysis should be a family of models or a specific model, and also to what we are trying to assess: the truth of model hypotheses, or their predictive accuracy. It is beyond the scope of this article to enter into this debate. We note simply that, even if an alternative (frequentist) approach to confirmation of base models is taken, the legitimacy of both double-counting I and II holds. Evidence used for calibrating base models is also used for determining their relative standing, or, in other words, for confirmation (see, for instance, Hitchcock and Sober [2004]).

Section 4 presents two analyses from the climate literature which exemplify the two special cases of double-counting II. The aim here is to show that climate scientists do engage in double-counting, even if they do not acknowledge it as such.

4 Climate Science Examples: Comparative Confirmation in Practice

There is considerable discussion in climate science about calibrating aerosol forcing. To give some background: Aerosols are small particles in the atmosphere. They vary widely in size and chemical composition and arise, for example, from industrial processes. Aerosols alter the Earth's radiation balance, and the aerosol forcing measures the extent that anthropogenic aerosols alter this balance. Anthropogenic aerosols influence the climate in two ways. First, they reflect and scatter solar and infrared radiation in the atmosphere (measured by the direct aerosol forcing). Second, they change the properties of

clouds and ice (measured by the indirect aerosol forcing). Overall aerosols are believed to exert a cooling effect on the climate.

The uncertainty about the magnitude of the aerosol forcing, in particular about the indirect aerosol forcing, is huge because little is known about the physical and chemical principles of how aerosols change the properties of clouds and ice, and how they scatter radiation. Consequently, it is standard practice to calibrate the aerosol forcing against data, and the aerosol forcing constitutes a prime example of calibration in climate science.

We will now show that in climate papers about the aerosol forcing we can find the two special cases of double counting II.

4.1 Confirmation due to better and worse best fits

The first paper we look at is Harvey and Kaufmann ([2002]). They compare the adequacy of two climate models (with model error) for simulating the observed warming of the past two and a half centuries. The two base models (derived from an energy balance model coupled to a two-dimensional ocean model) are:[16]

***M*1**: model instances that consider both natural and anthropogenic forcings to describe climate change (plus model error).

***M*2**: model instances that consider only anthropogenic forcings to describe climate change (plus model error).

They assume that the model error is such that none of the base-model hypotheses can be falsified by the data but where, roughly, the closer the simulations are to the observations, the better.[17] The evidence regarded as relevant for assessing the adequacy of the base models are the past record of mean surface temperature changes, interhemispheric surface temperature changes, surface temperature changes in the northern hemisphere, and surface temperature changes in the southern hemisphere. This evidence is used to simultaneously calibrate the aerosol forcing and the climate sensitivity. (The climate sensitivity measures the mean temperature change resulting from a doubling of the concentration of carbon dioxide in the atmosphere.) Motivated by physical considerations, the initial ranges considered are $[0, -3]$ for the aerosol forcing and $[1, 5]$ for the climate sensitivity.

They proceed as follows: among all the model instances of $M1$ and $M2$, Harvey and Kaufmann identify a model instance which best matches the data.

[16] The base model $M1$ ($M2$) does not consist of one model to which different forcing values can be assigned. It consists of several different models, which consider different anthropogenic and natural influences (different anthropogenic influences), to which different forcing values can be assigned. Hence Harvey and Kaufmann compare two sets of models.

[17] They do not assume any observation error.

Then they apply a statistical test to determine whether other model instances differ significantly from the best instance. In this way they arrive at a set of best performing models instances. (Denote this set by MB and let MB^C be the model instances of $M1$ and $M2$ which are not in MB.) It turns out that MB only includes instances of $M1$. Consequently, they conclude that there is confirmation: $M1$ (natural and anthropogenic forcings) is more adequate for simulating the past temperature record than $M2$ (only anthropogenic forcings). Furthermore, they use the same data to calibrate the aerosol forcing: the instances of $M1$ in MB correspond to an aerosol forcing range of $(-1.5, 0]$, which is thus regarded as the likely range.

Harvey and Kaufmann can be seen as engaging in double-counting II. Their procedure can (roughly) be reconstructed in Bayesian terms, as per Section 4. The model error is probabilistic.[18] Furthermore, because initially they are indifferent about the exact forcing values, they assume a uniform prior over the aerosol forcing and climate sensitivity conditional on $M1$ and $M2$.[19] Their procedure comes close to assigning to the probability of the data a much smaller value, given MB^C, than to the probability of the data, given MB. (That is, $Pr(E|MB^C)/Pr(E|MB)$ is much smaller than 1, for example, $\frac{1}{9}$.) Then, because MB only includes instances of $M1$, it follows that the probability of the data given $M1$ is much higher than the probability of the data given $M2$. Consequently, probabilistic confirmation theory yields that $M1$ is confirmed relative to $M2$ and that the aerosol forcing is very likely in the range $(-1.5, 0]$.

To conclude, Harvey and Kaufmann justifiably use the same data for calibration and comparative confirmation. They engage in case (1) of double counting II, i.e. there is confirmation because the average fit with the evidence is better for $M1$ than for $M2$. Note that we are not here assessing other aspects of the experimental design. For instance, climate scientists may debate the relevance of the past ocean temperature change data for comparing the models' adequacy. As stressed earlier, that is a different question not to be confused with double-counting.

4.2 Confirmation due to more and less plausible forcings values

As a second case let us compare the models of Knutti *et al.* ([2002]) and Knutti *et al.* ([2003]). Knutti *et al.*'s ([2002], [2003]) concern is to construct models which are adequate for long-term predictions of temperature changes (within the error bounds) until 2100 under two important emission scenarios. They

[18] Their method implies that (roughly) the smaller the model error, the better, and that none of the models can be falsified. However, apart from this, the assumptions about the model error remain unclear. It would be desirable to spell out these assumptions because this is needed for specifying the models' adequacy.

[19] Likewise, we assume that each of the different models in $M1$ ($M2$) are equiprobable (see Section 4.1).

assume that the model error is discrete (cf. Section 3). The two base models (derived from a dynamical ocean model coupled to an energy- and moisture-balance model of the atmosphere) are:

> ***M*1**: model instances considered by Knutti *et al.* ([2002]). There are five different ocean setups and the carbon cycle is not accounted for explicitly (the carbon cycle determines how emissions are converted into concentrations in the atmosphere).[20]
>
> ***M*2**: model instances considered by Knutti *et al.* ([2003]). There are ten different ocean model setups and the carbon cycle and its uncertainty are explicitly accounted for with a parameterization.[21]

The evidence that they regard as relevant for assessing the adequacy of these models is past mean surface temperature changes and ocean temperature changes.

All the elements needed to compare the two base-model hypotheses in the framework of probabilistic confirmation theory are present in Knutti *et al.* ([2002], [2003]). The evidence is used to calibrate simultaneously the indirect aerosol forcing and the climate sensitivity. Motivated by physical estimates, Knutti *et al.* ([2002], [2003]) assume that, conditional on $M1$ and $M2$, the indirect aerosol forcing is initially normally distributed with the mean at -1 and a standard deviation of 1.[22] The climate sensitivity is assumed to be initially uniformly distributed over [1,10], conditional on $M1$ and $M2$.

Knutti *et al.* ([2002], [2003]) then calculate the posterior probabilities for model instances, i.e. the likelihood of an arbitrary model-hypothesis instance given the data, assuming that $M1$ ($M2$) is true. A model-hypothesis instance is regarded as consistent if the average difference between the actual and the simulated observations is smaller than a constant.[23] A posterior probability is zero for inconsistent model-hypothesis instances. Consistent model-hypothesis instances are assigned a probability proportional to the prior probability over the forcings values (i.e. over the model instances[24]). It turns out that a posterior probability distribution over the forcings is the same for $M1$ and $M2$, implying the indirect aerosol forcing is likely (with approximate

[20] The ocean setups of $M1$ and $M2$ differ: the ten ocean setups of $M2$ do not include the five ocean setups of $M1$.

[21] Because of the different ocean setups, the base model $M1$ ($M2$) does not consist of one model to which different forcing values can be assigned but of five (ten) different models to which different forcing values can be assigned. Hence the sets of models $M1$ and $M2$ are compared.

[22] They also discuss the case of a uniformly distributed aerosol forcing. However, the case of the normal distribution will be more insightful here.

[23] The constant equals the standard deviation of the model ensemble, which in climate science is regarded as a measure of model error. They also assume that there is observation error. To account for it, the difference of the observed and modelled temperature is divided by the uncertainty of the observed warming (Knutti *et al.* [2002], [2003]).

[24] Knutti *et al.* ([2002], [2003]) assume that each of the five (ten) different models constituting the base model class $M1$ ($M2$) are equiprobable (cf. Section 4.2).

probability 0.90) to be in the range [−1.5, 0.2). In short, the consistent model instances of $M1$ span the same range of forcing values as the consistent model instances of $M2$. As all consistent model instances are regarded as having equivalent fit with the data (because postulated model error is discrete), we conclude that there is no comparative confirmation.

Now suppose that for $M1$ a posterior probability distribution over the forcings would have been different, say, the likely (with probability 0.90) aerosol forcing range would have been [−2.7, −1]. Then the data would have been justifiably used both for calibration and comparative confirmation of the base-model hypotheses. This would have been an example of case 2 of double counting II. $M2$ would have been confirmed relative to $M1$ because the specific instances of $M2$ favoured by the evidence are more plausible than the specific instances of $M1$ favoured by the evidence.

5 Old Evidence

We have seen that double-counting is not illegitimate, at least by Bayesian confirmation standards, and is, moreover, practised by some climate scientists. This problematizes assertions that double-counting is clearly bad. The remainder of the article considers reasons why double-counting may yet be, for the most part, inapplicable in the climate-model context. Note that the reasons we canvas concern the failure of calibration and/or confirmation of base models. Nothing we say in these final sections supports the position that separate data should be used for calibration and confirmation.

We start with what seems a prevalent concern: the evidence in question was used to formulate the climate-model hypotheses, and so is old evidence that is not suitable for further confirmation purposes. This appears to be a concern of Stainforth *et al.* ([2007a]):

> Development and improvement of long time-scale processes are therefore reliant solely on tests of internal consistency and physical understanding of the processes involved, guided by information on past climatic states deduced from proxy data. Such data are inapplicable for calibration or confirmation as they are in-sample, having guided the development process.

The term 'in-sample' is ambiguous here: on the one hand, it apparently refers to evidence belonging to a different time(or spatial) period from the predictions of interest (we discuss this issue in subsequent sections), yet on the other hand it seems to refer to old evidence, i.e. evidence already taken into account in model development. Since these two issues come apart,[25] they deserve separate treatment.

[25] Consider: it is possible to find 'new' evidence from the same time period as the 'old' evidence.

Our current concern is updating on old evidence. How might this problem manifest? It helps to consider a paradigm case: Imagine that a detective announces that the most plausible hypothesis, given the expensive earring and strands of hair found at the crime scene, is that the rich Lady visiting the manor killed the host. Clearly the evidence has already been taken into account in announcing that this hypothesis is the most plausible one. In Bayesian terms, the current plausibility of the hypothesis—the relatively high probability Pr(rich-Lady hypothesis)—is already in effect a posterior probability, given the evidence. It would thus be a mistake to further confirm the rich-Lady hypothesis with respect to the same evidence. One can still assess the confirmatory power of the old evidence, but this requires constructing a 'counterfactual' probabilistic belief function, Pr', representing what the detective's beliefs would have been, if the evidence, E, was not already known. It follows that $Pr'(E|\text{rich-Lady hypothesis}) < 1$. One may also appeal to the counterfactual prior probability of the rich-Lady hypothesis, Pr'(rich-Lady hypothesis).[26]

To better appreciate the problem, it is helpful to consider the overall confirmation from two independent pieces of evidence, say E_1 and E_2, according to Bayes' theorem. In such case, the overall confirmation of, say, H_1 relative to H_2 depends on the product of the two likelihood ratios:

$$\frac{Pr(E_1|H_1)}{Pr(E_1|H_2)} \times \frac{Pr(E_2|H_1)}{Pr(E_2|H_2)}. \tag{6}$$

It would be a mistake, of course, to treat the one piece of evidence, E, as if it were two pieces of independent evidence, and thus take confirmation due to E as:

$$\frac{Pr(E|H_1)}{Pr(E|H_2)} \times \frac{Pr(E|H_1)}{Pr(E|H_2)}. \tag{7}$$

This is what it means to update again on old evidence, or use the same evidence two times over for confirmation. It is effectively what would happen if, say, our detective further confirmed the rich-Lady hypothesis with respect to the same crime-scene data, and concluded that it was even more plausible that she was the murderer.

Let us now return to climate models. The way we have characterized calibration in Section 3 already guards against this old-evidence updating, to some extent. As mentioned, the problem set-up is crucial to a defensible Bayesian analysis. When calibrating and comparing two base-model

[26] Admittedly, it is not clear how to construct such a 'counterfactual' probabilistic belief function, and the controversy about its interpretation runs deep. This is not our present problem, however, and we note that others offer ways to make sense of these counterfactual probabilities (see, for instance, Eells and Fitelson [2000]).

hypotheses, we must assign all the specific instances of these models appropriate conditional priors, i.e. probabilities that do not yet take the evidence into account. Then the evidence can be used to calibrate or discriminate further between the model instances (and between the base models too, as per double-counting II). This is effectively the procedure that is followed in the case studies of Section 4: suitable conditional prior probabilities are initially selected and then updated in light of the temperature data.

Of course, evidence might be unwittingly used two times over for calibration and/or confirmation. Indeed, Frame *et al.* ([2007]) note this danger in the context of assessing climate models. They caution against calibrating and/or confirming twice with the same evidence, not realizing that the evidence already informed the conditional prior probability distributions over instances of the base models. In short, updating on old evidence is problematic, and practitioners should be careful to avoid doing this. But this is not an inevitable problem, and the remedy is not to use separate data for calibration and confirmation. The remedy is simply not to calibrate and confirm model hypotheses two times over with the same evidence.

There may be a lingering concern that prior probabilities for the base-model hypotheses themselves already incorporate the evidence, especially if base models with additional forcings or parameters are constructed expressly to achieve better fit with the data. So the base-model hypotheses are only a subset of the full space of possible models, and hence assigning each an equal prior probability would be to over-estimate their initial plausibility. The situation seems analogous to the murder case above—the base models that climate scientists work with are considered plausible precisely because the evidence has already been taken into account in selecting them. Just as the murder detective does not bother to mention various people near the crime scene who may have been under greater suspicion if the evidence were otherwise, climate scientists have presumably already dismissed a large number of possible base models in favour of the few under consideration that seem to have the potential to permit a reasonable fit with the evidence. It would then seem wrong to use the evidence a second time over for confirmation.

Notwithstanding this concern, we can still calibrate and assess comparative (incremental) confirmation in terms of a counterfactual probabilistic belief function where $Pr'(E|H_i) < 1$, indicating that the evidence, E, is not already known.[27] Furthermore, as mentioned above, even if the base-model hypotheses are only a subset of the full space of model hypotheses—the ones deemed most plausible in light of the evidence—one can still estimate counterfactual

[27] In the rest of the article, we return to assessing confirmation in terms of the regular probabilistic belief function Pr. Where there is a problem of old evidence, however, this should be understood as standing in for the appropriate counterfactual probability function, which we denote here as Pr'.

prior probabilities for the base-model hypotheses. Presumably, the counterfactual prior probabilities for these base models should not add to 1, but to some probability <1. Determining the appropriate probability mass to assign to the set of base-model hypotheses may be quite tricky. But this problem affects only non-comparative and, ultimately, absolute confirmation, and where we want to assess how confident we should be, overall, in our models. And again, this has nothing to do with double-counting. In any case, the assessment of non-comparative and absolute confirmation of climate models is plagued with even bigger difficulties, and we will get to these in Section 7.

For now we continue to analyse why even calibration and comparative confirmation may fail in the climate-model context. In particular, we turn now to concerns about the (ir)relevance of past data.

6 Doubts about the Relevance of Past Data

There is an important difference between the climate studies discussed in Sections 4.1 and 4.2. In the Harvey and Kaufmann study, past data were used to calibrate/confirm base-model hypotheses concerning past climate behaviour, whereas in the Knutti *et al.* studies, past data were used to calibrate/confirm base-model hypotheses concerning long-term future climate behaviour (policy makers are most interested in this long-term future climate behaviour). The latter is more controversial than the former and, as we will see in this and the next section, may be what some climate scientists have in mind when they make negative comments about calibration and confirmation. This section discusses whether particular past data are relevant for assessing the adequacy of climate-model hypotheses in predicting future climate variables of interest. The next section will discuss the concern that climate models are based on assumptions that may not hold in the future, and hence there is considerable uncertainty about the full space of models that are possibly adequate for predicting future climate.

Let us initially confine our analysis to the model instances of a single base-model hypothesis, for example, L (Equation (1) in Section 3). Assume that the model hypotheses, denoted $L_{1,1}, L_{1,2} \ldots$, this time concern whether the line in question (plus probabilistic model error) accurately predicts $y(t)$ for future times $t \geq t*$. Our question here is: Can past data, i.e. data for $t < t*$, help in calibrating L?

The answer: it all depends on what is the implicit relationship between $t < t*$ and $t \geq t*$, i.e. the implicit extension of the model instances of L that span $t \geq t*$ into the past. One possibility is that the past values depend strongly on the future values and vice versa, a special case being where each line in L for $t \geq t*$ is associated with just one and the same line for $t < t*$. In this case, past data (past values for $y(t)$), E, are clearly relevant for comparing

$L_{1,1}, L_{1,2} \ldots$.[28] The likelihood ratios $Pr(E|L_{i,j})/Pr(E|L_{k,l})$ may be calculated as before.[29]

Another possibility, of course, is that the past values are independent of the future values, a special case being where each line in L for $t \geq t*$ is associated with any line for $t < t*$. That is, each line hypothesis in L, such as $L_{1,1}$, is implicitly associated with a whole set of extended models:[30]

$$y(t) = \begin{cases} t + 1 + N(0, \sigma) & \text{if } t \geq t*; \\ \gamma t + \theta + N(0, \sigma) & \text{if } t < t*. \end{cases} \tag{8}$$

Here E, i.e. past values for $y(t)$, will be irrelevant for comparing instances of L, the reason being that all instances of L are associated with the same pasts, and so E does not distinguish these instances. That is to say that the pertinent likelihoods for calibration—$Pr(E|L_{i,j})/Pr(E|L_{k,l})$—all equal 1. So in this case there is no calibration of L and thus, in a sense, no double-counting I.

The analysis of double-counting II is essentially the same. In this case, we are comparing two base-model hypotheses, for example, L and Q (Equations (1) and (3) in Section 3) where the concern is whether the models accurately predict $y(t)$ for future times $t \geq t*$. Consider the special case where every model instance of L or Q is implicitly extended into the past in the same variety of ways.[31] In this case, past data, E, again does not favour any instance of either model over any other instance of either model, and we obtain $Pr(E|L)/Pr(E|Q) = 1$. Neither base hypothesis is confirmed relative to the other. So in a sense there is no double-counting II (in addition to no calibration and no double-counting I). Of course, this is just a special case. If the values of past and future variables were dependent, past data may confirm one base-model hypothesis over another.

This scenario of independence is what some climate scientists seem to have in mind when they say:

> Statements about future climate relate to a never before experienced state of the system; thus, it is impossible to either calibrate the model for the forecast regime of interest or confirm the usefulness of the forecasting process. (Stainforth *et al.* [2007a], p. 2146)

We have here the grounds for a charitable interpretation of climate scientists' claim that data cannot be used to calibrate and confirm climate models. As

[28] Note that the various frequentist estimators used in model selection, such as the Akaike estimator, assume an unchanging physical reality or data generation process.

[29] Recall our earlier Section 3.1, which notes that the likelihoods are more precisely stated $Pr(E|L_{i,j}$ and $B)$, and so on, where B is background knowledge. Here background knowledge about the implicit relationship between past and future is very important for determining the value of the likelihood.

[30] Also, the implicit conditional probabilities for the past extensions are assumed not to vary for the $L_{i,j}$.

[31] Again, the implicit conditional probabilities of the extensions are assumed not to vary for the $L_{i,j}$ and $Q_{i,j}$.

suggested by the quote, one might say that calibration is impossible when the future climate variables in question (or the equations that adequately predict them) are considered independent of the past data at hand (or the equations that adequately predict them).[32] It is important to note that the extent to which the point applies in climate science is controversial. Some climate scientists suggest that the future values of prominent climate variables, including precipitation and even average global temperature rise, are more or less unconstrained by the past values of these or other variables (for example, Frame *et al.* [2007]; Stainforth *et al.* [2007a]). Other climate scientists apparently do not think it so plausible that past values for at least some prominent climate variables are irrelevant to their future values (for example, Knutti *et al.* [2002], [2003]; Randall and Wood [2007]). In any case, the claim that calibration fails and there is no confirmation of model instances or model hypotheses in a particular context is very different from the claim that double-counting is 'bad practice'. Moreover, using separate past data for calibration and confirmation is no remedy for this problem.

7 Non-comparative Confirmation and Catch-alls

We have thus far been concerned with confirmation of one model hypothesis relative to another. Yet certain statements from climate scientists concerning calibration suggest that what is at issue is whether the evidence confirms the predictions of a model *tout court*, i.e. relative to its complement (non-comparative confirmation). We first show that double-counting is also legitimate for non-comparative confirmation. Then we explain why, nonetheless, confidence in future climate predictions may be hard to amass. The difficulties arise when climate models are based on assumptions which are suspected to be wrong in the future. Again, the problem cannot be solved by employing separate data for calibration and confirmation.

In some cases, assessing non-comparative confirmation is relatively straightforward. The relevant likelihood ratio involves a model (a base model or a specific instance) and its entire complement. For instance, the degree to which evidence, E, confirms base model hypothesis, M, relative to its entire complement is (where $N, \ldots, Z$ are the mutually exclusive base model hypotheses that exhaust the complement of M):

$$\frac{Pr(E|M)}{Pr(E|\neg M)} = \frac{Pr(E|M)}{Pr(E|N) \times Pr(N|\neg M) + \ldots + Pr(E|Z) \times Pr(Z|\neg M)}. \quad (9)$$

[32] A case which often arises in climate science is that the equations for adequately predicting the past and future climate variables are considered identical in form, yet the parameters in these equations have values for past and future that are independent.

As before, this likelihood ratio may be greater than, less than, or equal to 1, corresponding to M being confirmed, disconfirmed, or neither, relative to its complement.

Here again it must be noted that the final probability of M, i.e. $Pr(M|E)$, is a further matter, and depends also on the prior probability $Pr(M)$. This section too focuses just on the extent to which evidence incrementally confirms or raises confidence in a model, this time relative to its complement. An examination of the above expression reveals, however, that non-comparative confirmation nonetheless requires substantial information regarding the prior probabilities of base models, in the form of conditional probabilities like $Pr(N|\neg M)$. So the comments at the end of Section 5 regarding difficulties in estimating the prior probabilities of base models are pertinent here.

Further problems arise when the full set of base models under consideration is believed not to be exhaustive, and yet we are unable to specify what is missing (there are 'known unknowns'). In other words, we have a range of plausible base-model hypotheses plus a catch-all, i.e. a hypothesis to the effect 'none of the above is true'. One can easily see that non-comparative confirmation in these conditions is difficult to assess. The relevant likelihood is (where M is a base-model hypothesis, and hypotheses N, ... together with the catch-all C exhaust the complement of M):

$$\frac{Pr(E|M)}{Pr(E|\neg M)} = \frac{Pr(E|M)}{Pr(E|N) \times Pr(N|\neg M) + \ldots + Pr(E|C) \times Pr(C|\neg M)}. \quad (10)$$

The problem is that the likelihood associated with the catch-all, $Pr(E|C)$, let alone the probability $Pr(C|\neg M)$, is very difficult to evaluate. How do we estimate the probability of some evidence conditional on the truth of a hypothesis which we cannot actually specify?

The common sentiment in climate science seems to be that there is indeed a catch-all, especially when the models' purpose is to predict future climate. Nonetheless, some studies appear to proceed under the assumption that model hypotheses may be confirmed (or disconfirmed) to some degree in non-comparative terms, given evidence. Most plausibly, in these cases the catch-all is either negligible, or else it is not completely unspecified, and some climate scientists think they know enough about it to at least have rough estimates for $Pr(E|C)$. If at least a rough estimate for $Pr(E|C)$ can be given (as well as rough estimates for all other terms in the expression above), the main conclusions drawn about double-counting and comparative confirmation carry over. In particular, double counting II is legitimate for non-comparative confirmation and can arise for two reasons (cf. Section 3): (1) better fit of the model or the complement of the model with the evidence and/or (2) the specific instances of the model that are favoured by the evidence

may be more plausible or less plausible than the instances of the complement favoured by the evidence.

So far so good, but some climate scientists do not think the prospects for non-comparative confirmation of model hypotheses concerning the future are so rosy. First, note that if past data is considered independent of the future (cf. Section 6), there cannot be non-comparative confirmation because there is no confirmation of one base-model hypothesis relative to another, or indeed the catch-all.

Second, even if past data are relevant, many scientists worry that climate models (which are based on our understanding of climate processes to date) invoke assumptions which may not hold in the future.[33] Consider:

> For these processes, and therefore for climate forecasting, there is no possibility of a true cycle of improvement and confirmation, the problem is always one of extrapolation and the life cycle of a model is significantly less than the lead time of interest. (Stainforth *et al.* [2007a], p. 2147)

One might interpret this view as follows: if base-model hypotheses concern future predictions, then the catch-all is overwhelming. Future climate behaviour may differ from that of the past/present in unanticipated ways, and so we are unable to specify even roughly the appropriate likelihoods of the relevant catch-all.

At this point it should be mentioned that climate models are designed to accurately simulate mean surface temperature changes. They fail to simulate absolute mean surface temperatures to a similar level of accuracy. In particular, the simulated mean surface temperature changes are derived from simulated surface temperature values that show biases of several degrees Celsius on many regions of the Earth. The same holds for other variables such as ocean temperatures (Knutti *et al.* [2010]; Randall and Wood [2007], p. 608 and supplementary material). There is nothing in principle wrong with modelling temperature changes rather than absolute temperatures. When one variable is too difficult to predict, often scientists succeed instead in predicting a simpler variable such as an average or a change in that variable. However, many climate scientists argue that the reason why climate models fail to accurately simulate absolute temperatures is because important processes are ignored which may become relevant for adequately predicting long-term future climate behaviour of interest (for example, Stainforth *et al.* [2007a]). From this, doubts arise about whether current climate models will adequately describe the relevant aspects of the future climate.

[33] Note that while these two concerns are logically distinct, they are of course closely related in the climate context. This is because the scientific reasons for doubting the relevance of past climate data have much overlap with the reasons for positing significant uncertainty about the future.

Climate scientists seem to take different views on the extent of our uncertainty about the future. But in the case of radical uncertainty, non-comparative confirmation of any one, or the whole set, of our climate-model hypotheses concerning the future is indeterminate, even if past data are relevant for comparing pairs of hypotheses. Overall confidence in any single model or the full set of models cannot increase.[34] This position regarding non-comparative confirmation is reflected in the following statement concerning the modelling of future climate:

> We take climate ensembles exploring model uncertainty as potentially providing a lower bound on the maximum range of uncertainty and thus a non-discountable [unable-to-be-ignored] climate change envelope [range of climate-change predictions]. (Stainforth *et al.* [2007b], p. 2167)

We now turn to an example in climate science which highlights the controversies surrounding the relevance of past data and the overall adequacy of climate models for future predictions.

8 Climate Science Example: Non-comparative Confirmation and Catch-Alls in Practice

Our example for non-comparative confirmation with a catch-all is Knutti *et al.* ([2003]), already discussed in Section 4.2, and again concerns the aerosol forcing. Recall that Knutti *et al.* aim to construct models which are adequate for long-term predictions of the temperature changes until 2100 under two emission scenarios (within the error bounds), and that the model error is discrete. The two base models are:

$\boldsymbol{M}$: Models instances of Knutti *et al.* ([2003]).

$\boldsymbol{C}$: Catch-all.

Recall that mean surface temperature changes and the ocean warming are regarded as relevant to assess the adequacy of the models, and they are used to constrain the indirect aerosol forcing and the climate sensitivity. Motivated by physical estimates, for the aerosol forcing a uniform distribution over $[-2, 0]$ is chosen conditional on M or C.[35] For the climate sensitivity, a uniform distribution over $[1, 10]$ is chosen conditional on M or C.

The data are used for calibration: Knutti *et al.* ([2003]) calculate the likelihood of an arbitrary model-hypothesis instance given the data, assuming that M is true. Because of the uniform prior distribution over the forcings values, consistent model-hypothesis instances are equiprobable given the data;

[34] Moreover, applying full Bayesian reasoning, the posterior probabilities of the climate-model hypotheses would also be indeterminate due to the indeterminate likelihood ratios. Most plausibly, in the case of a radically unspecified catch-all, the prior probabilities would be indeterminate as well.

[35] Knutti *et al.* ([2003]) also discuss the case of a normally distributed aerosol forcing. See Section 4.2.

inconsistent model-hypothesis instances have zero probability (a model-hypothesis instance is regarded as consistent if the average difference between the actual and the simulated observations is smaller than a constant). The conclusion is that the likely range (summing to probability 0.93) of the indirect aerosol forcing is [−1.2,0). Furthermore, Knutti *et al.* seem to claim that the data confirm M relative to the catch-all because the fit with the data is very good and the model could have (easily) failed to simulate the data.

As already discussed in Section 4.2, Knutti *et al.* ([2003]) use elements of probabilistic confirmation theory. However, when reconstructing this as a case of non-comparative confirmation, what is missing are the values of $Pr(E|M)$ and, in particular, of $Pr(E|C)$. The crucial question is whether $Pr(E|M)/Pr(E|C) > 1$. If it is, then probabilistic confirmation theory will yield that the data are justifiably used for non-comparative confirmation and calibration. There will be double-counting II for reason (1)—the model instances of M provide a better fit with the data than the catch-all.

It should come as no surprise that the answer to this question is controversial. Knutti *et al.* ([2003]) tend to an affirmative answer. They seem to claim that confidence in the future predictions of M has increased. However, if Stainforth *et al.* ([2007a]) are right that past data are not relevant to the future climate predictions of interest (as discussed in Section 6) or that the probabilities associated with the catch-all cannot be precisely specified (as discussed in Section 7), then the answer will be negative. The data simply will not confirm M relative to the catch-all.

The fact that there is controversy among climate scientists about such fundamental and policy-relevant questions highlights the need to think more carefully about them. Whatever the outcome, this controversy is not about the problem of double-counting.

9 Concluding Remarks

The main contribution of this article is the untangling and clarification of worries concerning double-counting. We have argued that the common position—that double-counting is bad and that separate data must be used for calibration and confirmation of base-model hypotheses—is by no means obviously true. This is not to say there are no other fundamental concerns about the confirmatory power of evidence or about uncertainty in climate science. It is crucial, however, that the various issues are articulated and distinguished, if we are to make progress in assessing confidence in climate models and their predictions.

Our claim is that double-counting, in the sense of using evidence for calibration and confirmation, is justified by at least one major approach to confirmation—the Bayesian or relative likelihood approach. Calibration of a base-model hypothesis is all about determining which specific instances of

the base model are confirmed relative to other specific instances. We call this double-counting I. Furthermore, we showed that, according to Bayesian standards, the same evidence may be used for calibration and for incrementally confirming one base-model hypothesis relative to another, or relative to its entire complement. We call this double-counting II. For both double-counting I and II, calibration and confirmation can be seen as the same process in the sense that evidence is used to do both calibration and confirmation simultaneously. We appealed to studies in climate science to show that these two forms of double-counting are in fact practised by some climate scientists, even if they are not acknowledged as such.

In the latter parts of the article, we acknowledged and discussed important worries about calibration and confirmation in the climate-modelling context that may be marring the double-counting debate. In some cases, evidence already informs the prior assessment of model instances. If so, it cannot be used again for calibration and confirmation—this would be using the same evidence two times over. More fundamentally, there is often controversy about what evidence is relevant to whether a model achieves its purpose. Treating irrelevant evidence as if it was relevant and using this evidence for confirmation or calibration is also bad practice. Indeed, some climate scientists state strongly that future climate variables of interest are more or less unconstrained by the available past climate data. The upshot is that this past climate data are irrelevant for assessing the adequacy of models for predicting the future. Hence, there can be no calibration or double-counting. A related but subtly different concern is that climate models are based on assumptions which may not be applicable in the future. This would imply that one cannot hope to even roughly determine the likelihood of the catch-all hypothesis with respect to adequately predicting the future, and so non-comparative confirmation, let alone absolute confirmation, would be indeterminate.

We noted that climate scientists disagree about whether these worries are all justified. In any case, the worries concern whether data are useless for confirmation and/or calibration. Problems of this kind cannot be remedied by using separate data for calibration and confirmation. We thus suggest that practitioners be clearer about their targets. Suspicions about the legitimacy of double-counting should not be confused with other important issues, such as what evidence is relevant for confirmation given the modelling context at hand, whether issues of old evidence are appropriately handled, or whether the worry is justified that climate models are based on assumptions which will not hold in the future.

Acknowledgements

Earlier versions of this article have been presented at the third conference of the European Philosophy of Science Association, the 2010/2011 London

School of Economics Discussion Group Meetings on Climate Science and Decision-Making, the 2011 Bristol Workshop on Philosophical Issues in Climate Science, the First Annual Ghent Metaphysics, Methodology, and Science Program, the 2011 Geneva Workshop on Causation and Confirmation, the 2011 Stockholm Workshop on Preferences and Decisions, and the 2012 Popper Seminar. We would like to thank the audiences for valuable discussions. We also thank Reto Knutti, Wendy Parker, and David Stainforth for their helpful comments.

Katie Steele
Department of Philosophy, Logic and Scientific Method
London School of Economics and Political Science
Houghton Street
London WC2A 2AE, UK
k.s.steele@lse.ac.uk

Charlotte Werndl
Department of Philosophy, Logic and Scientific Method
London School of Economics and Political Science
Houghton Street
London WC2A 2AE, UK
c.s.werndl@lse.ac.uk

References

Anderson, T. L., Charlson, R. J., Schwartz, S. E., Knutti, R., Boucher, O., Rodhe, H. and Heintzenberg, J. [2003]: 'Climate Forcing by Aerosols: A Hazy Picture', *Science*, **300**, pp. 1103–4.

Burnham, K. P. and Anderson, D. R. [1998]: *Model Selection and Multimodal Inference*, Berlin and New York: Springer.

Eells, E. and Fitelson, B. [2000]: 'Measuring Confirmation and Evidence', *Journal of Philosophy*, **97**, pp. 663–72.

Forster, M. and Sober, E. [1994]: 'How to Tell When Simpler, More Unified, or Less *Ad Hoc* Hypotheses Will Provide More Accurate Predictions', *British Journal for the Philosophy of Science*, **45**, pp. 1–35.

Frame, D. J., Faull, N. E., Joshi, M. M. and Allen, M. R. [2007]: 'Probabilistic Climate Forecasts and Inductive Problems', *Philosophical Transactions of the Royal Society A*, **365**, pp. 1971–92.

Harvey, D. and Kaufmann, R. K. [2002]: 'Simultaneously Constraining Climate Sensitivity and Aerosol Radiative Forcing', *Journal of Climate*, **15**, pp. 2837–61.

Hitchcock, C. R. and Sober, E. [2004]: 'Prediction versus Accommodation and the Risk of Overfitting', *British Journal for the Philosophy of Science*, **55**, pp. 1–34.

Howson, C. [1988]: 'Accommodation, Prediction, and Bayesian Confirmation Theory', *Proceedings of the Biennial Meeting of the Philosophy of Science Association*, 1988, pp. 381–92.

Knutti, R. [2008]: 'Should We Believe Model Predictions of Future Climate Change?', *Philosophical Transactions of the Royal Society A*, **366**, pp. 4647–64.

Knutti, R. [2010]: 'The End of Model Democracy: An Editorial Comment', *Climatic Change*, **102**, pp. 395–404.

Knutti, R., Stocker, T. F., Joos, F. and Plattner, G.-K. [2002]: 'Constraints on Radiative Forcing and Future Climate Change from Observations and Climate Model Ensembles', *Nature*, **416**, pp. 719–23.

Knutti, R., Stocker, T. F., Joos, F. and Plattner, G.-K. [2003]: 'Probabilistic Climate Change Projections Using Neural Networks', *Climate Dynamics*, **21**, pp. 257–72.

Knutti, R., Furrer, R., Tebaldi, C., Cermak, J. and Meehl, G. [2010]: 'Challenges in Combining Projections from Multiple Climate Models', *Journal of Climate*, **23**, pp. 2739–58.

Mayo, D. G. [2010]: 'An *Ad Hoc* Save of a Theory of Adhocness? Exchanges with John Worrall', in D. G. Mayo and A. Spanos (*eds*), *Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability, Objectivity and Rationality of Science*, Cambridge: Cambridge University Press, pp. 155–69.

Parker, W. S. [2010]: 'Comparative Process Tracing and Climate Change Fingerprints', *Philosophy of Science*, **77**, pp. 1083–95.

Parker, W. S. [2009]: 'Confirmation and Adequacy for Purpose in Climate Modelling', *Aristotelian Society Proceedings*, **83**, pp. 233–49.

Randall, D. A. and Wood, R. A. [2007]: 'Climate Models and Their Evaluation', in S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K. B. Averyt, M. Tignor and H. L. Miller (*eds*), *Climate Change 2007*: *The Scientific Basis*, Cambridge: Cambridge University Press, pp. 589–662.

Rodhe, H., Charlson, R. J. and Anderson, T. L. [2000]: 'Avoiding Circular Logic in Climate Modeling', *Climatic Change*, **44**, pp. 419–22.

Shackley, S., Young, P., Parkinson, S. and Wynne, B. [1998]: 'Uncertainty, Complexity, and Concepts of Good Science in Climate Change Modelling: Are GCMs the Best Tools?', *Climatic Change*, **38**, pp. 159–205.

Tebaldi, C. and Knutti, R. [2007]: 'The Use of the Multi-Model Ensemble in Probabilistic Climate Projections', *Philosophical Transactions of the Royal Society A*, **365**, pp. 2053–75.

Stainforth, D. A., Allen, M. R., Tredger, E. R. and Smith, L. A. [2007a]: 'Confidence, Uncertainty, and Decision-Support Relevance in Climate Predictions', *Philosophical Transactions of the Royal Society A*, **365**, pp. 2145–61.

Stainforth, D. A., Downing, T. E., Washington, M., Lopez, A. and New, M. [2007b]: 'Issues in the Interpretation of Climate Model Ensembles to Inform Decisions', *Philosophical Transactions of the Royal Society A*, **365**, pp. 2163–77.

Worrall, J. [2010]: 'Error, Tests, and Theory Confirmation', in D. G. Mayo and A. Spanos (*eds*), *Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability, and the Objectivity and Rationality of Science*, Cambridge: Cambridge University Press, pp. 125–54.