Predicting under Structural Uncertainty: Why not all Hawkmoths are Ugly

Karim Bschir[1] and Lydia Braunack-Mayer[2]

[1]Department of Humanities, Social and Political Sciences, ETH Zurich

[2]Department of Mathematics, ETH Zurich

November 4, 2018

1 Introduction

Predicting the future behavior of complex dynamical systems with the help of nonlinear models is an important part of scientific practice. However, making predictions from nonlinear models is often affected by severe uncertainties. In recent years, there has been an extensive debate about the epistemic limitations of model-based predictions not only in the philosophy of science, but also within the scientific community.[1]

[1] See for instance a recent special issue of *Science* on prediction and the limits of predictability in current science (Jasny and Stone, 2017).

In an important contribution to this debate, Frigg et al. (2014) have made the claim that uncertainty about the true dynamical structure of a nonlinear model seriously debilitates our ability to make decision-relevant predictions.[2] In analogy to the well-known Butterfly Effect, Frigg et al. introduce the Hawkmoth Effect, which arises from a sensitive dependence on structural model error (SME). Just as the Butterfly Effect describes the limitations that initial condition uncertainty imposes on the predictive power of nonlinear models, the Hawkmoth Effect highlights the supposedly disastrous consequences that small errors in a model's structure can have for making predictions from that model. However, the authors say little about the scope of their argument. In certain passages, they seem to claim that our current modeling practices offer no effective countermeasures against the Hawkmoth Effect. More specifically, they think that ensemble modeling approaches provide no remedy against SME. In this paper, we challenge this claim. We argue that Frigg et al. ignore the numerous tools of modern statistics for the handling of model uncertainty, and thus overstate the epistemic consequences of SME in general. We do concede, however, that their argument points at serious limitations in the context of climate science. These limitations arise due to specific properties of climate models and the ensembles used in climate science. Ultimately, our contribution aims at continuing Frigg et al.'s train of thought by investigating the scope of their argument from the perspective of modeling practice.

We begin, in the next section, by explaining the three steps of the argument by which Frigg et al. support their claims. The first step, to which the majority of their paper is devoted, involves the presentation of an example that illustrates the consequences of SME for prediction. The example shows that the probability distribution derived from a

[2] See also Bradley et al. (2014), Frigg et al. (013a), Frigg et al. (013b).

model with only a small structural error can differ drastically from the probability distribution that would have been obtained using the true dynamical equations. This difference is the result of the Hawkmoth Effect. In the second step, Frigg et al. generalize their example. To this end, they refer to a couple of mathematical theorems in order to show that structural stability cannot be taken for granted in most nonlinear models. In a third and final step, they claim that there is currently no method that can mitigate against the Hawkmoth Effect, and that no such method is likely to be developed. As a consequence of this, they argue that the burden of proof should be shifted. The default assumption should be that nonlinear models are structurally unstable and, hence, susceptible to the Hawkmoth Effect. Those who intend to use a nonlinear model for predictive purposes must prove that their model is structurally stable. If they fail to do so, probabilistic predictions derived from the model must be seen as genuinely unreliable.

Section three provides a discussion of Frigg et al.'s argument. We highlight the fact that a crucial part of the argument implicitly relies on the assumption that, in face of their limited knowledge of the true structure of a system, modelers rely on what the authors call the closeness-to-goodness link, i.e. the notion that a model whose structure is close enough to the true structure of system, will produce good enough predictions. We believe that this assumption—that modelers rely on closeness-to-goodness—is not generally true. In many domains where models are used for prediction, closeness-to-goodness is rarely seen as a sufficient criterion for deciding that predictions derived from a nonlinear model are likely to be "good enough". We show that even if there is a strong correlation between nonlinearity and structural instability there are effective strategies against the impact of SME that are currently applied in scientific practice. In fact, there are a number of approaches that can guard against the potential consequences of SME. We present three of these approaches: Bayesian Model Averaging (BMA), bootstrap model averaging, and the Super Learner. All these approaches operate under the rationale that making a prediction from an ensemble of models offers the best protection against SME. While Frigg et al. anticipate this objection, we explain why their reasons for doing so are unfounded by highlighting relevant insights from the statistical literature. We also briefly discuss examples of model averaging in scientific practice. The first two examples, from wildlife ecology and microbiology, illustrate the use of BMA to handle and deal with uncertainties about model structure. The third and fourth examples demonstrate the applicability of bootstrap model averaging in studies of equipment degradation and of the association between mortality and air pollution. Our aim is to explore cases where the Hawkmoth effect can be brought under control, thereby showing that there are domains where a constructive solution is available. Consequently, the severity of the Hawkmoth effect depends on the domain one is looking at. The problem is hard in some domains, while it can be efficiently dealt with in other cases.

Section four focuses on the discussion of SME in the context of climate modeling. We show that Frigg et al.'s argument has special relevance in climate science. In climate science, the complexity and size of climate models makes them less accessible to ensemble approaches, and the number of models used in an ensemble has to be small due to limitations in computational power.

We conclude that Frigg et al. have overstated the epistemic consequences of SME *in general*. While SME is an important source of uncertainty that needs careful attention from practicing scientists, it does not always debilitate our ability to make informative predictions of dynamical systems. In many contexts, commonly used modeling practices can help to identify structural errors and well-established statistical methods allow scientists to take measures against the impact of model uncertainty. The mere existence of such tools indicates the awareness that scientists have of the implications of structural uncertainties. However, limitations in computational power make these tools not equally applicable in all disciplines. Frigg et al.'s skeptical conclusion needs to be taken seriously in domains where models are large and ensembles are small. Climate science is such a domain.

2 The Hawkmoth Effect

Frigg et al. introduce the Hawkmoth Effect by analogy to the well-known Butterfly Effect. The Butterfly Effect describes the limitations that small errors in the initial conditions of nonlinear models impose on their predictive power.[3] Similarly, the Hawkmoth Effect describes the consequences that small errors in the structure of a nonlinear model can have for predictions derived from that model. The effect arises not because of initial condition error, but rather because of a model's sensitive dependence on structural model error (SME).

The main point of Frigg et al.'s analysis is to show that the Hawkmoth Effect seriously debilitates our ability to make decision-relevant predictions from nonlinear dynamical models. We can mitigate against the Butterfly Effect by applying a probability distribution to the set of initial conditions. Instead of evolving one point in the state space, the dynamics of the model can then be used to evolve a probability distribution of initial conditions. The Butterfly Effect then tells you that the longer you project into the future, the more your distribution will spread out. Or, in other words,

[3] See Lorenz (1972).

system sates that are initially close together in the state space can move far apart under nonlinear dynamics. The Hawkmoth Effect cannot be dealt with in the same way, because SME does not lead to a spreading of a probability distribution of states over time, but to an incorrect evolution of the probability distribution itself. A model with even a small SME can produce probabilities for outcome states that differ significantly from the true probabilities. This means that, if a nonlinear model has only the slightest SME, then, according to Frigg. et al., its ability to produce decision-relevant probabilities is lost entirely.[4]

For the purpose of illustration, Frigg et al. invite their readers to imagine two variations of Laplace's all-knowing demon: the freshman apprentice and the senior apprentice. In contrast the demon himself, who has precise knowledge of the initial conditions and the exact dynamical equations of any given system, the apprentices fall short of their master's absolute epistemic capacities. The senior apprentice shares all the capacities of the demon with the exception of observational omniscience; she has only incomplete and imprecise information about initial conditions. This limits her predictive power. Because of the Butterfly Effect, according to which arbitrarily small errors in the initial conditions can lead to significantly different future states, the senior apprentice is not able to make precise point predictions. She mitigates against the Butterfly Effect by making probabilistic forecasts. She puts a probability distribution over his initial conditions in order to account for his uncertainty regarding the latter. She then uses the dynamical equations to evolve the distribution in time. Her result is a probabilistic prediction that tells her the future state of his system up to a certain degree of precision.

[4] Readers familiar with Frigg et al.'s article may prefer to proceed to the next section at this point.

The freshman apprentice, in addition to imperfect knowledge of initial conditions, also suffers from imperfect dynamical knowledge. In other words, the equations of his model are affected by SME. This means that the equations at his disposal are structurally different from the equations of the demon, which describe the real dynamics of the system (Frigg et al., 2014, p. 35). It is important to note that what Frigg et al. have in mind here is not simple parameter error, where the parameters in the model are such that they do not correctly align with the true values. SME refers directly to a difference in the mathematical structure of the equations due to missing or superfluous nonlinear terms. In order to compensate for his ignorance, the freshman apprentice makes an assumption that allegedly allows him to use his model for predictive purposes despite its structural deficiencies. He assumes the existence of the closeness-to-goodness link: if the model is close enough to the true equations, than it will accurately predict the system's future behavior. In order to quantify closeness to the true equations and goodness of predictions, Frigg et al. introduce two metrics. Closeness to the true equations is measured in terms of maximal one-step-error, and goodness of prediction is measured in terms of the relative entropy between the true and the predicted probability distribution. If the difference in entropy between the two distributions is small enough then the prediction can be seen as accurate (Frigg et al., 2014, pp. 35-36).

The goal of their demon-apprentice thought experiment is to make the situation in which real-world modelers find themselves in explicit. Like the freshmen apprentice, in many practical tasks modelers are confronted with both initial condition uncertainty and SME. But while initial condition uncertainty can be dealt with by the use of probabilistic methods, the limitations imposed by SME cannot be handled by applying the closeness-to-goodness link. The simple reason for this, as Frigg et al. argue, is that the closeness-to-goodness link does not hold in general. The upshot of their argument is to show that arbitrarily small errors in the structure of the dynamical equations of a nonlinear model lead to arbitrarily big errors in the predictions derived from those equations. Although the structure of the equations used in the model might differ from the true equations only to a small degree, and although the error introduced in each step of the prediction can be arbitrarily small, the difference between the true probability distribution and the distribution produced by the model after certain time, can turn out to be maximal. Hence, the closeness-to-goodness link does not hold.

Frigg et al.'s argument has the following structure. In a first step they present a case study of a generic nonlinear model and show that, for this example, the closeness-to-goodness link does not hold. Their example is a model of a fish population in a pond, the true equation for which is given by:

$$N_{t+1} = (1-\epsilon)4N_t(1-N_t) + \epsilon\frac{16}{5}(N_t(1-2N_t^2+N_t^3)). \tag{1}$$

Only the demon and his senior apprentice have access to this equation. The freshman apprentice tries to predict the fish population with the help of the logistic map, modeling the number of fish at a time point $t+1$ with the following equation:

$$N_{t+1} = 4N_t(1-N_t).^5 \tag{2}$$

[5] In their example, instead of the absolute number of fish, the authors use the ratio of the number of fish per cubic meter and the maximum number of fish that can be accommodated in one cubic meter. That number always lies in the interval between 0 and 1. See Frigg et al. (2014, p. 36).

This model has SME but when ϵ in equation (1) is small the graph of the true equation is almost indistinguishable from that of the model. This means that, according to the closeness-to-goodness link, the logistic map should predict the population of the fish pond in the future well enough. However, it turns out that after a certain point in time the probability distribution derived from the model differs significantly from the real distribution (see p. 38, Figure 2 in Frigg et al., 2014). What is more, before this point of failure the model seems to be well-behaved. Its predictions are close to the fish pond's true population. This is where the Hawkmoth Effect's ugliness comes from. The failure of the closeness-to-goodness link is not immediately visible, but kicks in abruptly at an indeterminable point in the future. The model's usability for predictive purposes is destroyed.

In the second step of their argument, the authors generalize the results suggested by their case study. To this end, they refer to mathematical theorems in order to show that the Hawkmoth Effect is a generic character of models that do not have a mathematical property of dynamic systems called structural stability. Roughly speaking, a dynamic system has structural stability if small changes in its equations have very little effect on the behavior of its trajectories. In other words, structural stability means that similar models with identical initial conditions will make similar predictions in the future. Their generalization shows that structural instability is not an idiosyncratic property of the logistic map and of their sample case, but that it applies to a very large class of nonlinear systems.

In order to justify the claim that structural instability is a generic phenomenon of differential equations, the authors refer to a mathematical theorem by Palis and Smale (1970), according to which a nonlinear flow is structurally stable if and only if it satisfies

Axiom A and the so-called strong transversality condition. Axiom A and the strong transversality condition are abstract properties of mathematical objects, the former requiring that a system is uniformly hyperbolic and the latter requiring that stable and unstable manifolds must intersect transversely at every point. Certain mathematical objects have been shown to have these properties: isomorphisms on smooth manifolds (Mañé, 1987) and flows (Hayashi, 1997). However, for a large number of nonlinear systems, the two conditions are not satisfied. Furthermore, Smale (1966) has shown that the set of structurally stable systems is open but not dense. Over all, these mathematical results suggest that the vast majority nonlinear systems lack the property of structural stability.[6]

In a third step, the authors tie these abstract findings back to the epistemic consequences drawn from their case study. The closeness-to-goodness link only holds if the true dynamics of a system are structurally stable. Only in this case can a model with an equation close to the truth lead to accurate predictions. And, since a large number of nonlinear systems are *not* structurally stable, the closeness-to-goodness link may not be presupposed by default. As a consequence, the authors demand that the burden of proof should be shifted. The default assumption should be that nonlinear models lack structural stability and that whoever intends to use a nonlinear model for predictive purposes must prove that their model is structurally stable. It is only in this case that the Hawkmoth Effect does not come into play. Otherwise, small errors in the structure of the model will lead to the catastrophic consequences that the authors describe in their case study.[7]

[6] All the mentioned papers are cited in Frigg et al. (2014, p. 47).

[7] We have to admit at this point, that the mathematics of these stability proofs are

In the following, we present a critique of Frigg et al.'s conclusions based on considerations concerning the handling of SME in scientific practice. Frigg et al. argue that this impact places a burden on modelers to prove that their models are structurally stable and, hence, do not fall prey to the consequences of SME. We do not contend that it is the responsibility of scientists to show that their predictions are not susceptible to the Hawkmoth Effect, or that the consequences of failing to do so can be critical. We do, however, contend with the idea that guaranteeing structural stability offers the only protection against the Hawkmoth Effect. We argue that even if there exists a strong correlation between nonlinearity and structural instability, there are efficient strategies against the impact of SME, which are actually applied in current scientific practice.

beyond our mathematical capacities and we are unable to judge their validity to its full extent. It is important to note, however, that one of the authors admitted in oral communication that the mathematics used in their claim to generality are intricate. Even for many theoretical physicists, the meaning of these abstract results for physical systems is apparently not straightforward (See Roman Frigg's talk at Center for Advanced Studies in Munich in November 2013. `https://itunes.apple.com/ch/podcast/chaos-beyond-butterfly-effect/id741597015?i=1000236728753&mt=2`, retrieved November 28, 2017). However, we not put into question the fact that the logistic map used in Frigg et. al.'s example, as well as most nonlinear models in current modelling practice fall under the scope of these theorems and that the theorems are relevant for the present context.

3 Handling Model Uncertainty in Practice

3.1 Overemphasizing closeness-to-goodness

Frigg et. al.'s mathematical considerations about structural stability lead to the conclusion that—in absence of a proof to the contrary—it must be assumed that a given set of nonlinear differential equations is structurally unstable. Hence the closeness-to-goodness link cannot be established and probabilistic predictions based on models containing nonlinear differential equations may fall prey to the Hawkmoth Effect. The epistemic conclusions Frigg et al. draw from their analysis are quite drastic. In their own words: "Many operational probability forecasts are therefore unreliable as a guide to rational action if interpreted as providing the probability of various outcomes" (Frigg et al., 2014, p. 57).

In Frigg et al.'s example case, the freshman apprentice represents the situation in which real-world modelers find themselves: modelers have imperfect knowledge of a system's true dynamics and cannot identify the true initial conditions, but they believe that they can make informative predictions with models that are close-to-good. The idea behind the closeness-to-goodness link is, to repeat, "the maxim that a model that is close enough to the truth will produce predictions that are close enough to what actually happens to be good enough for a certain predictive task" (Frigg et al., 2014, p. 35). The authors' analogy suggests that assuming closeness-to-goodness is the best that modelers can do, i.e. they choose models whose structure is close to the true dynamical structure of the system under scrutiny. However, real-world modelers often also consider other aspects of a model before they decide that its predictions are likely to be "good enough" for a certain predictive task.

Unlike Frigg et al.'s freshman apprentice, who only works with a single model, real world modelers frequently look to a comparison with different models for hints about the impact of minimal changes in the equations on the dynamics of the system. One way to asses the extent to which structural errors influence predictions is a procedure called a sensitivity analysis. A scientist performing a sensitivity analysis of her predictions of the fish population in the pond would compare the freshman apprentice's model with other potential models. She might do this quantitatively, with Monte Carlo methods or within a formal Bayesian framework. Or she might perform a simple ad hoc analysis of the impact that adding or removing structural components from the model has on her predictions. These approaches vary in their procedures but have the same underlying idea with respect to exploring model uncertainty: to critically examine and compare the predictions of models with different structures and to discover structural features of the model that have a critical influence on predictions. The standards and procedure for sensitivity analysis vary greatly between different fields. Saltelli et al. (2008) give an excellent general guide to sensitivity analysis.

A second criterion for goodness used in practice is the comparison of a model's predictions with observed data. A structural difference between a model and the real system's dynamics can translate to predictions that are visibly different from observed data. In many cases, the closeness-to-goodness link cannot be established precisely because the structure of the chosen model contains small errors, and calibration with observed data fails.

These practical considerations are important for scientists using models to predict dynamic systems. In Frigg et al.'s fish pond example, the freshman apprentice chooses a model without comparison with other models and ignoring any ways in which his chosen model's predictions might deviate from observed data. The freshman apprentice therefore constitutes an odd analogy for real-world modelers, because the latter have plenty of statistical tools at hand that allow them to either compare their preferred model's performance with other, structurally different models or to calibrate their model with past data. However, the feasibility applicability of such statistical tools as sensitivity analysis varies greatly between different fields. Despite the fact that in many relevant cases, structural errors can be spotted *a priori* by sensitivity analysis or preliminary calibrations, it remains true that they can not always be easily detected. In such cases, SME can have a serious impact on probabilistic predictions. Although Frigg et al. present their argument in a generic manner without a special focus on a specific field, and despite the fact that they formulate their epistemic conclusion in general terms, it his hard to overlook the fact that their intended domain clearly is climate science. In climate science, SME could indeed be a serious problem for the reasons laid out by the authors, and, as we will discuss later, for specific circumstances that limit the applicability of ensemble modeling approaches in climate science. These limitation are mainly related to the complexity and size of climate models and of the generally small number of models used in climate model ensembles. Other fields, where predictive models are also large and and cannot be easily calibrated with past data, might suffer from similar problems. In any case, this raises the question about the scope of Frigg et al.'s argument: For which kinds of modeling task does it pose a serious threat to probabilistic predictions, and what are the circumstances under which SME becomes a tamable problem?

Frigg et al. have already been accused of overgeneralizing their case and it has been noted that the scope of their argument is unclear. Goodwin and Winsberg (2016) distinguish between between a broad, provocative and a narrow, modest interpretation of Frigg et al.'s argument. Under the broad interpretation the predictive power of all nonlinear models is impaired by the Hawkmoth Effect. In this reading, Frigg et al.'s argument would have the most severe implications for most scientific modeling endeavors. In the modest interpretation, their considerations only extend to high resolution climate predictions as they are made, as for instance in the UKCP09 modeling project (for details see Section ... below). Goodwin and Winsberg go on to show that, presupposing the modest interpretation, Frigg et al.'s base case involving the logistic map generalizes to applications in climate science only along one of four possible dimensions. Frigg et al.'s generalization merely considers, according to Goodwin and Winsberg, the time evolution of the model. It does not consider the timescale, the way in which probabilities are generated from the model, or the kind of prediction one hopes to obtain from the model. Goodwin and Winsberg then conclude that Frigg et al.'s argument from analogy fails, because the set of real-world modeling projects that are relevantly similar to Frigg et al.'s base case on all four dimensions is small. They allegedly show that the sort of probabilistic statements generated in the UKCP09 are significantly different from the way probabilities are obtained in Frigg et al.'s base case. Furthermore, they point out that even if climate models are structurally unstable, it remains unclear what the relevant timescales would be for the instabilities to manifest themselves (Goodwin and Winsberg, 2016, p. 1128).

We have to admit at this point that we find Goodwin and Winsberg's objections not particularly enlightening, and we suspect that they might be misinterpreting the dialectical situation that Frigg et al.'s argument creates. First, it remains unclear to what extent probabilities created in real-world modeling projects are "significantly different" from the probabilities in Frigg et al.'s fish pond example. Second, Frigg et al. present the Hawkmoth Effect as a generic phenomenon, which arises as a consequence of certain mathematical properties of nonlinear models. If the Hawkmoth Effect is real, and if most nonlinear models are indeed unstable, the fact that the timescale of the effect is not well understood becomes secondary. It makes things even worse. If we know that the predictions produced by our best climate models can fail completely at some point in the future, but we have no clues as to when that point may be, our trust in those predictions gets even more jeopardized. Therefore, our criticism in the next section, in contrast to Goodwin and Winsberg's, will not such much directed against Frigg et al.'s arguments *per se*. Rather, it goes against their skepticism towards the use of ensemble methods for the handling of structural uncertainty. Thus the correct way to asses the scope of Frigg et al.'s arguments would be to ask if and under what conditions ensemble methods provide an effective countermeasure against the Hawkmoth Effect.

In recent years, modelers have gained access to a suite of statistical tools that can mitigate against the impact of structural errors. In the following sections we take a look at three methods that can be applied to guard against the impact of SME: Bayesian model averaging, bootstrap model averaging, and the Super Learner.

3.2 Bayesian Model Averaging

Model averaging is the general term for a method that accounts for model uncertainty by combining the results of multiple models. A basic presupposition of model averaging rests in the acknowledgement that we usually do not know the true structure of the system to be modeled, i.e. that we are indeed in the position of Frigg et al.'s freshman apprentice. But unlike Frigg et al.'s freshman apprentice, users of model averaging do not operate under the assumption that there is a single, correct model. Instead, they derive predictions from an ensemble of models without invoking or expecting that any one model in the ensemble represents the real system truthfully. In may cases, the acknowledgment of structural uncertainty is the main motivation for using ensemble methods. One of the most well-established and generic types of model averaging is Bayesian Model Averaging (BMA). In recent years, BMA has been increasingly used for the handling of model uncertainty in a wide variety of fields (see Clyde and George, 2004, p. 82). Drawing on Hoeting et al.'s *Bayesian Model Averaging: A Tutorial* (1999), we give a brief, generic description of the rationale behind BMA. We also describe two applications of BMA in current scientific practice.

Suppose that we are interested in using data, denoted by D, to predict the state of a system in the future, denoted by x. For a chosen ensemble M of models $M_1, ..., M_K$, the probability of x given the data D can be written as a weighted average of the probabilities under each of the models considered:

$$pr(x \mid D) = \sum_{k=1}^{K} pr(x \mid M_k, D) pr(M_k \mid D). \tag{3}$$

This quantity is called the state's posterior distribution. Another important quantity, the posterior probability for each model M_k in our chosen ensemble, is given by

$$pr(M_k \mid D) = \frac{pr(D|M_k)pr(M_k)}{(\sum_{l=1}^{K} pr(D \mid M_l)pr(M_l)}, \tag{4}$$

where $pr(M_k)$ is the prior probability that M_k is the true model, and

$$pr(D \mid M_k) = \int pr(D \mid \Theta_k, M_k) pr(\Theta_k \mid M_k) d\Theta_k \tag{5}$$

is the integrated likelihood of the model M_k with parameters Θ_k. $pr(\Theta_k \mid M_k)$ is the prior density of Θ_k given model M_k, and $pr(D \mid \Theta_k, M_k)$ is the likelihood of the data under the model M_k. This quantity depends critically on the choice of prior probabilities $pr(M_k)$, which should suitable for the modeling task at hand. Often the prior probabilities are taken from a distribution that represents the modeler's knowledge about the system. Importantly, each of these mathematical expressions is conditional on the chosen ensemble of models M. Using these quantities, we can make a prediction $\hat{x}$ of the system's future state x, given the data D and model ensemble M, by taking a weighted average:

$$\hat{x} = \sum_{k=1}^{K} \hat{x}_k pr(M_k \mid D), \tag{6}$$

where $\hat{x}_k$ is the prediction of model M_k. The weights are given by the posterior probability of each model. The rationale behind this approach is that obtaining predictions in this way gives better average predictive ability than making a prediction from any single model in the chosen ensemble.

BMA is widely used in many fields of research to account for model uncertainty. We highlight two recent examples, the first of which is from the field of microbiology. Shankar et al. (2015) study the effects of broad band antibiotic treatment on the gastrointestinal tract, in particular on fugal colonization. Their model captures the responses in microbiome community and host immune response to antibiotic treatment. BMA is applied to three separate model ensembles, one made up of different logistic regression models and two consisting of different linear regression models. For each ensemble, the authors apply BMA with a prior distribution called the spike-and-slab prior. Aside from accounting for model uncertainty, BMA helps the researchers to examine large numbers of variable configurations and to compute effect sizes efficiently. They use Markov-Chain Monte Carlo methods to explore a space of 10000 likely models.

Artelle et al. (2016) apply BMA in the context of wildlife ecology. The goal of their study is to test three hypotheses regarding the causes of human-grizzly conflict. They model the spatial and temporal variation in ecological predictors of patterns of conflict (fluctuations in the availability of salmon, limited food supply, problem individuals, regional population saturation). BMA is applied to predict the numbers of bears killed in specific geographical regions. Predictions are made from two different ensembles, each consisting of negative binomial regression models whose predictor variables are unique combinations of salmon biomass, spring and summer temperature, spring and summer precipitation, the total number of bears killed by hunters, and the total number of bears killed in attacks on humans. Each regression model is used to predict the number of grizzly bears killed in conflict. These predictions are then averaged according to the BMA procedure.

We choose examples from two distinct domains of research in order to highlight the fact that the feasibility of BMA approaches is highly dependent on the context. The specific problems that arise in different contexts may vary greatly in kind and severity. Typical issues that have to be addressed in modelling tasks of the above kind are: time-scale (i.e. how far in the future do predictions go and how fast does predictability decrease with time), weighting of models in the ensemble, choice of priors and, most importantly, the number of structurally different models available. In both of the above examples the models are relatively simple in terms of the causal structures that they have to represent. This makes it easy to create large ensembles (up to 10000 in the microbiome example), and BMA methods become more efficient when the space of structurally different models is large. In domains where model construction is expensive and models cannot be easily multiplied the effectiveness of ensemble methods as a remedy against SME is limited.

3.3 Frigg et al.'s objections against BMA

Frigg et al. offer a number of reasons for rejecting a Bayesian strategy against the consequences of SME. The first objection is practical. The authors believe that "it is unfeasible to generate predictions with an entire class of models." (Frigg et al., 2014, p. 56). Aside from the fact that it remains unclear what they exactly mean by "unfeasible" in this context, it certainly must be admitted that applying BMA can be computationally challenging. Making BMA predictions involves averaging estimates from what can be a large number of models, and involves calculating or approximating the often intractable likelihood of each model in the ensemble. As Montgomery and Nyhan note, "[t]hese computational difficulties led many early researchers to adopt simplifying assumptions and techniques that made BMA analyses more tractable but required significant trade-offs" (Montgomery and Nyhan, 2010, p. 249). However, recent theoretical and computational advances in Bayesian analysis have improved our ability to apply BMA in practice: "The combination of increased computing power, the development of more analytically tractable prior specifications, and the distribution of the BMA and Bayesian adaptive sampling (BAS) packages for [the statistical software] R have made these techniques far more accessible" (Montgomery and Nyhan, 2010, p. 249). These advances have made BMA computationally accessible, and give us reason to reject Frigg et al.'s generic complaint about the unfeasibility of ensemble approaches. Furthermore, the wide use of multi-model approaches in current scientific practice can be considered reason enough to suppose that the approach is indeed feasible in some specific cases.

The second objection is that "it is not clear how to circumscribe the relevant model class. This class would contain all possible models of a target system. But the phrase all models masks the fact that mathematically this class is not defined, and indeed it is not clear whether it is definable at all." (Frigg et al., 2014, p. 56). We agree. But it is important to note that good BMA predictions can be obtained without considering an entire class of models. For example, Markov Chain Monte Carlo (MCMC) methods can be applied to approximate the posterior distribution of a class of models from a representative sample of models. As Fernandez et al. demonstrate, MCMC methods used in conjunction with BMA can lead to good predictive results without a huge computational burden when the number of models under consideration is as large as 2.2 trillion (Fernández et al., 2001, p. 564). But also small, carefully chosen ensembles can give accurate predictions. Madigan and Raftery (1994) have shown that applying BMA with a small number of models chosen by Occams Window, a heuristic greedy-search algorithm for a subset of models with relatively high posterior probabilities, provides better predictive ability than making predictions from any plausible, single model. These results indicate that ensemble approaches can be powerful even if the ensemble used does not contain all possible models of the system. To be sure, the question if a chosen subset is representative of the entire class of possible models is a difficult one, and, again, highly context dependent. In domains where we have a few structurally homogenous models at hand (e.g. climate science) we have to be mindful of the possibility that the set of available models might not be representative of the class of all possible models. This does, however, not speak against the feasibility of BMA methods per se, but rather at the fact that operating with ensembles with structurally diverse models is desirable if BMA methods are used as a countermeasure against the effects of SME.

The third objection is a technical problem related to the choice of an appropriate measure of uncertainty on the class of models. More precisely, "the relevant class of models would be a class of functions, and function spaces do not come equipped with measures. In fact, it is not clear how to put a measure on function spaces" (Frigg et al., 2014, pp. 56). In the context of applying BMA in practice, this means that choosing model priors that adequately reflect our uncertainty about the true model can be, at best, difficult and, at worst, impossible. We admit that this is an open and well-recognized problem in both BMA and in Bayesian frameworks more generally. Ley and Steel (2009) have, for example, applied BMA to growth data with different priors and found that, in some cases, prior choice critically affected the posterior probabilities. However, this difficulty offers no reason for the general claim that BMA cannot be used to handle structural uncertainty. Modeling a nonlinear dynamic system is a practice that involves making a number of choices that are technically and theoretically difficult, and that critically affect the model's predictions. Choosing appropriate priors is no different, and modelers can refer to the literature or to other experienced modelers for help in making this critical decision. In practice, many users of BMA use default priors that are commonly used in their field of study, or they assume that all models are equally likely and apply vague priors, such as a uniform distribution. Others approach the problem directly and derive an uncertainty measure appropriate to their particular task, based on expert opinions or on the characteristics of their problem (Fragoso and Neto, 2015). Furthermore, as Gelman points out in reference to the general problem of choosing a prior in Bayesian methods, BMA-users can evaluate the effect of their choice of prior on their predictions: "[I]n practice one can check the dependence on prior distributions by a sensitivity analysis: comparing posterior inferences under different reasonable choices of prior distribution" (Gelman, 2002, 1634). Instead of forcing us to reject BMA as a strategy against structural uncertainty, the difficulty Frigg et al. identify simply highlights the need for carefully considered choices when applying Bayesian averaging methods.

The fourth objection points to the fact that "we, like the Freshman, are restricted to sampling from the set of all conceivable models, which need not contain a perfect model even if such a thing exists." (Frigg et al., 2014, p. 57). This means that even if we have addressed the first three problems, it may be impossible to find an ensemble that contains the true model. However, using an ensemble can be helpful even when all models in the ensemble are wrong. A number of studies have shown that BMA improves predictions even when we do not know if our ensemble contains the true model (Madigan and Raftery 1994; Montgomery and Nyhan 2010; Fragoso and Neto 2015). Averaging over the predictions of merely approximately correct nonlinear models is likely to be beneficial even when none of the models is perfect. (Wasserman, 2000, p. 103). Practicing scientists rarely aim to find a true model. Instead, they look for models that are able to achieve a particular task. For example, in epidemiology models are often used to predict outbreaks of common pathogens. The standard deterministic SIR (Susceptible-Infectious-Recovered) model divides a population into three distinct groups, those susceptible to a disease, those who can infect others with the disease and those who have recovered and are immune. An individual's transition from group to group is governed by a series of nonlinear, differential equations. Such a model vastly oversimplifies the true dynamics of the spread of an infectious disease. Yet, models such as the SIR model can approximate a system's true dynamics accurately enough to be helpful in a wide array of applications.[8] Thus multi-model approaches can lead to reliable predictions even in cases where ensembles contain only approximately true models, i.e. models that only approximately represent the true structure of the system. To be sure, whether we can know if the used models are approximately true depends again on considerations that are highly dependent on context and the causal makeup of the system under scrutiny.

These considerations show that BMA a is widely applicable and well-established strategy for handling structural uncertainties. While its application is not without challenges, addressing these challenges is a focus of ongoing research efforts. What is more, BMA is just one of a growing number of ensemble-based modeling methods that offer strategies against the consequences of SME.

3.4 Beyond Bayesian Model Averaging

Here we present two, equally plausible methods for handling the consequences of SME: bootstrap model averaging and the Super Learner. These methods may be preferable for some modeling tasks and offer an alternative for modelers who would rather avoid the complexities of a Bayesian approach. Together with BMA, these and other model

[8] Brauer (2008) gives an excellent introduction to the use of compartmental models in epidemiology.

averaging methods provide the practicing scientist with a tool kit for guarding against SME.

Bootstrap model averaging is a simple method that helps to handle structural uncertainty. First introduced twenty years ago by Buckland et al. (1997), it approaches model averaging from a frequentist perspective. Like BMA, bootstrap model averaging starts with the choice of an appropriate ensemble of models. But unlike BMA, it does not depend on a choice of prior probabilities. Instead, it makes use of data simulated by bootstrap sampling, a procedure that generates a dataset by randomly sampling with replacement from the original dataset. Given a chosen model ensemble, bootstrap model averaging proceeds in the following way:

1. Create a bootstrap sample: randomly sample data with replacement until the sample has as many elements as the original dataset.

2. Fit each model in the ensemble to the bootstrap sample.

3. Use a quantitative criteria, such as the Akaike Information Criteria (AIC), to select the model that best fits the bootstrap sample.

The scientist repeats this procedure a large number of times, assigning weights to each model by counting the number of times that the model was 'best'. Predictions are made by averaging the weighted predictions of each model in the chosen ensemble. Since its introduction in the 1990s, this basic procedure has been advanced and adapted to suit different modeling contexts and has been shown to outperform BMA in some contexts (see Martin and Roberts 2006 and Roberts and Martin 2010).

A number of recent studies demonstrate bootstrap model averaging's ability to address the problem of SME in real-world contexts. Baraldi et al. (2013) show that bootstrap model averaging can improve predictions of the remaining useful lifetime of degrading technical equipment. The authors simulate data from degrading turbine blades operating at high temperatures, aiming at assessing whether bootstrap model averaging can help to predict a particular kind of damage called creep growth. They find that bootstrap approaches can reliably predict the remaining useful lifetime of degrading equipment and, more importantly can accurately quantify the uncertainty of such predictions.

Martin and Roberts (2006) show that bootstrap model averaging can reduce structural model uncertainty in time series studies of the relationship between air pollution and mortality. A common feature of using time series to explore the association between air pollution and mortality "is that myriad modeling choices must be made to arrive at an 'optimal' model" and "[t]he procedure of selecting a single 'best' model may ignore the model uncertainty, which is inherently involved in searching through the set of candidate models to determine the best one" (Roberts and Martin, 2010, p. 131). They assess the performance of a bootstrap procedure on simulated data, finding that it was more accurate than prediction from a single model (Martin and Roberts, 2006). Notably, the bootstrap procedure performed well even when implausible models were included in the ensemble. They also found that a more complex version of bootstrap model averaging, involving a second layer of bootstrap sampling, performs better than predicting from a single model, a BMA procedure and the single bootstrap procedure (Roberts and Martin, 2010). These results show that bootstrap model averaging is a viable method for handling the consequences of SME in practical contexts.

Another, more recent ensemble method from machine learning, the Super Learner, offers yet another data-driven strategy against the consequences of SME. Introduced in

2007 by van der Laan et al. (2007) and later adapted for prediction (Polley and van der Laan, 2010; Polley et al., 2017), this method is motivated by the use of a procedure called cross validation. In cross validation, a dataset is split into two distinct subsets called a test set and a training set. The training set is used to fit the model and the test set is used to evaluate the model's ability to accurately predict new data. A slightly more complex variation of this procedure, V-fold cross validation, is often used to select a best model from a number of candidates. Basically, the Super Learner is a model averaging procedure that cleverly incorporates V-fold cross validation into its choice of weights. A scientist who applies the Super Learner proceeds as follows:

1. Split the dataset into V equally sized groups.
2. Create V training sets by leaving each of the V subsets out of the dataset. The v^{th} test dataset is the v^{th} subset. The v^{th} training dataset is the union of all remaining subsets.
3. Fit each model to the V training datasets.
4. Use each of the V model fits to predict the corresponding test datasets.

Ultimately, this procedure is used to select the set of model weights that minimize the cross validated risk, a quantity that estimates each model's accuracy to out-of-sample data. Roughly speaking, these weights reflect how accurately models in the ensemble would predict new data. Finally, a Super Learner prediction is made exactly as in BMA and bootstrap model averaging, by taking a weighted average of predictions from each model in the ensemble. But this method does not involve a choice of prior and, interestingly, does not require that the true model be in the ensemble. Theory shows that the Super Learner's special way of choosing the weights ensures that, asymptotically, the resulting prediction is as good as the prediction of any single model in the ensemble. In the case that the true model is in the chosen ensemble, this means that the Super Learner guarantees accurate predictions. In the case that the ensemble does not contain the true model, the Super Learner prediction will be the best available (van der Laan et al., 2007).

BMA, bootstrap model averaging and the Super Learner are only three of a number of model averaging methods that can be applied to guard against structural uncertainties. We have focused on BMA because of its wide treatment in both theoretical and applied science. The bootstrap approach offers an equally viable alternative for those who prefer to avoid Bayesian methods. The Super Learner is a new method from machine learning that, while not widely applied in practice, is gaining traction in the scientific community and has a theoretical grounding that guarantees good results. While each of these methods is designed for modeling problems involving observed data, the general principles behind model averaging can be applied in problems that do not involve fitting a model to observations. In sum, model averaging does provide an efficient and practical tool for mitigating against the impact of SME on predictions of nonlinear dynamical systems in many relevant contexts.

4 Structural Uncertainty in Climate Science

As mentioned earlier, it is important to highlight the fact that the intended domain of Frigg et al.'s considerations is climate science. Climate science commonly operates with large and complex models. This means that if ensemble methods are to be used for prediction in climate sciences the computational burden is usually huge and the size of the ensembles is generally quite small. However, a major advantage of the above discussed statistical tools was precisely their ability to handle *large* ensembles of *simple* models. So the question arises whether our optimism towards ensemble modeling approaches as a feasible countermeasure against SME was premature. Maybe our objection that the closeness-to-goodness link is rarely applied in real science only holds in in fields where scientists operate with simple and highly idealized models, i.e. in cases where they know that their models are unlikely to be truthful representations of the real system. But in fields where models are intended to be comprehensive and realistic representations of a complex physical system, where the computational cost of creating models is high, and where they cannot be easily calibrated with past data, closeness-to-goodness is often the only applicable criterion to evaluate the predictive skill of a model. This would mean that Frigg et al.'s conclusions about the limitations of our predictive practices imposed by SME are correct in fields that deal with large realistic models—such as climate science—, but are exaggerated in domains where models are idealized and simple—such as in the above mentioned examples from wild life ecology, engineering or environmental science. Let us thus take a brief look at another publication, where the authors apply their argument to a real modeling project in climate science.

In their 2013 paper, Frigg, Smith and Stainforth investigate the impact of the Hawkmoth Effect on the UKCP09 climate modeling project.[9] The UKCP09 is a highly localized model that provides high-resolution ensemble forecasts of climate during the twenty-first century in the United Kingdom. The information provided by the UKCP09

[9] See also Frigg et al. (2015).

is supposed to guide decision-making and mitigation measures that would help prevent the impacts of climate change in the UK.

Frigg et al. (2013b) identify two assumptions made by the UKCP09: The proxy assumption and the informativeness assumption. The proxy assumption holds that a multi model ensemble can serve as a viable proxy for the real world, and that the effects of structural errors of a given model can be assessed by comparing it to the ensemble. The informativeness assumption is that a given model is informative about the real world and that there exists a measure (the so-called discrepancy term) for the discrepancy between the model and the real world (Frigg et al., 013b, pp. 892-893).[10] The informativeness assumption is rejected by the authors based on their claims about the impact of SME (Frigg et al., 013b, p. 894). A model can only be seen as informative if the inference from closeness to truth to the reliability of the model outputs could be drawn in general. But, as the argument in Frigg et al. (2014) shows, the closeness-to-goodness link only holds under certain conditions, and not in general. The proxy assumption is rejected based on a frequent general objection against ensemble modeling that model ensembles only help to attenuate the impact of structural errors in cases where the the models in the ensemble are independent. If the models in the ensemble share a common structural error the ensemble as a whole suffers from the same deficiencies as a single model with SME. It might turn out that the distribution produced by the ensemble is far away from the target. Therefore the proxy assumption is unjustified. A multi model ensemble provides no reliable measure for the distance of a

[10] To be precise, in Frigg et al.'s presentation, the informativeness assumption is part of the core assumption, which also contains an assumptions about the discrepancy term. For reasons of simplicity and relevance, we only focus on the informative assumption here.

given model to the real system (Frigg et al., 013b, p. 895. See also Smith, 2002). Applied to actual modeling projects such as the UKCP09 this amounts to the attribution of complete failure: "[T]he aim of UKCP09 was to provide trustworthy forecasts now, and this, we have argued, they fail to do" (Frigg et al., 013b, p. 896).

The UKCP09 thus serves as a paradigm case where the abstract mathematical considerations about structural stability have immediate practical implications, insofar as they shed skepticism on the usability of nonlinear models for predictive purposes on highly localized scales. It has been argued elsewhere that the lack of predictive capacity on localized scales prohibits effective adaptation measures because we simply do not know what to adapt to. As a consequence, claims to the effect that adaption to climate change is more feasible and cost effective than mitigation of green house gases turn out to be ill founded (Oreskes et al., 2010). All this shows that a lot depends on the validity of Frigg et al.'s arguments. If they are correct, this could have immediate practical consequences for climate policy. Therefore, it is of crucial importance to understand the scope of their argument correctly.

Recall that Frigg et al,'s main skepticism against ensemble approaches in climate science consisted in the rejection of the proxy assumption, according to which multi model averages can serve as a viable proxies for the real world. They believe that the proxy assumption fails if there are common errors in the ensemble. And we have good reasons to suspect that our current climate model ensembles do contain such common errors (Parker 2011, cited in Frigg et al. 2013 [p. 895]; see also Knutti et al., 2010), in particular if the number of models in an ensemble is small.

The problem of common errors in ensemble modeling is widely acknowledged in the literature (Sexton et al., 2012, p. 2516; Parker, 2011, both cited in Frigg et al., 013b; see also Parker, 2014). So there is little doubt that model dependence and common biases constitute serious problems for every ensemble modeling effort. But it is also important to note that we observe a tendency towards the use of structurally less uniform sets of models. A higher degree of diversity might make the interpretation and combination of the models more difficult. “On the other hand”, as Tebaldi and Knutti (2007) note, “they [less uniform ensembles] will probably sample a wider range of structural uncertainties in that case, and will be *reducing the concern about common biases*” (Tebaldi and Knutti, 2007, 2071).[11] The underlying rationale here is straightforward: The problem of common errors can be attenuated if we use less uniform ensembles, and since the uniformity of an ensemble is a function of its size, the larger the ensemble the more diverse it becomes.

General circulation models (GCMs) in climate science usually involve several components of the Earth system. The four traditional components of a GCM are atmosphere, land surface, ocean, and sea ice. In recent decades, there has been a trend to move from traditional GCMs to more complex earth system models (ESMs) which include further components such as land ice, biogeochemical cycles, aerosols, ecosystem models and more.[12] The tendency towards ever more complex models appears to be continuing. For example, there have been efforts to include even simulations of year-to-year variations in the emergence and loss of leaves by trees and other plants into climate models (Puma et al., 2013). The transition from the commonly used HadCM3 to the more recent state-of-the-art HadGEM1 model involved an improved representation of clouds, water vapor and radiative properties, increased horizontal and vertical

[11] Emphasis added.

[12] For an overview of the development of climate models see Figure 1.13 in Cubasch et al. (2013, p. 144).

resolutions, a new gravity wave scheme, as well as the inclusion of an interactive sulphur cycle among other things (Pope et al., 2007). Overall, we have observed a strong tendency towards more realistic models in the past, i.e. towards models that represent the full complexity of the climate system as comprehensively as possible.

Large and complex climate models can be highly useful to improve our understanding of the climate system. In light of Frigg et al.'s arguments we must, however, be highly cautions when using such models for predictive purposes. Although they might be extremely realistic representations of the reals climate, their predictions could be seriously flawed due to the Hawkmoth Effect, and ensemble modeling may not provide an effective remedy, because the complexity of the models only allows us to construct ensembles with few members. In small ensembles common structural errors are likely, and therefore predictions from such ensembles may not significantly better than predictions from a single model. This issue becomes even more pressing in the context of highly localized modeling projects such as the UKCP09 where the ensemble consisted of no more than 12 models.

The widespread efforts in the climate community to produce realistic models can in fact be interpreted as the application of what Frigg et al. call the closeness-to-goodness link. The effort of climate scientists to generate realistic models can indeed be seen as an indication for their implicit belief that models which are closer to the truth produce better predictions. And in these cases, Frigg et al.'s skeptical conclusion does in fact hold. Their argument shows that the closeness-to-goodness link fails *in gerneral* due to the lack of structural stability of nonlinear models.

The issue of model complexity in climate science is widely acknowledged. Running comprehensive ESMs comes with high computational cost. This allows only for a limited number of experiments on higher resolutions, which hinders the systematic exploration of uncertainties and studies of long-term evolutions of climate (Randall et al., 2007, p. 643). Due to the problems associated with complex models, the recent literature has produced a handful of contributions, which seem to go against the trend to use ever more realistic and complex climate models for predictions. To mention just one example, Seneviratne et al. (2016) introduce a scaling approach for the prediction of local regional extreme events. They show that temperature extremes and heavy precipitation events robustly scale with global temperatures and accordingly with cumulative CO_2 emissions in a range of scenarios. The basic idea is to use a simplified emulated version of the original models, in order to make the testing of hypotheses more simple and to render uncertainties more graspable. The structural uncertainties of high-resolution-models are recognized and circumvented by the use of the scaling approach. The scaling approach may be seen as an attempt to produce decision-relevant predictions on the regional scale that are not derived from the solution of structurally complex and highly realist nonlinear models, but which rely on a few very simple causal relationships such as the relationship between global temperature increase and cumulative CO_2 emissions.

Combining Frigg et al.'s argument with our considerations about some statistical tools for ensemble modeling, we may infer a tentative recommendation for future modeling efforts in climate science: If we could move towards simpler models containing few and simple causal relationships, and if these models prove to be useful for predictive purposes (as it happens to be the case in Seneviratne et al. 2016), this might render the above mentioned statistical tools for ensemble modeling more accessible to climate modeling, which in the end would indeed attenuate the serious problem of SME in climate science.

To be sure, the use of ensemble approaches in climate science will still pose other difficulties even if we will have simpler models at hand in the future. Finding ways to create sufficiently large ensembles with appropriate degrees of diversity is only one of many challenges. Another problem of interest is finding adequate weighing schemes for climate model ensembles. As mention earlier, model interdependence of models in an ensemble leads to the problem of common biases, which is more virulent in small ensembles (for studies on the lack of model independence see Bishop and Abramowitz, 2013 and Jun et al., 2008). Therefore, having a quantitative measure for the interdependence between models and ipso facto of the uniformity of an ensemble would be highly desirable. Until very recently, the prevailing approach for ensemble-based climate predictions has been model democracy, i.e. giving all models in the ensemble the same weight despite the fact that there often are good reasons to believe that certain models in an ensemble might outperform others for a given predictive task (Knutti, 2010; for multi-model approaches in climate science see e.g. Tebaldi et al., 2004 and Weigel et al., 2010). In a recent contribution to the debate, Knutti et al. (2017) have suggested a weighing scheme for multi-model projects that accounts for differences in model performance and model interdependence. The weighing scheme contains a term for the distance metrics between individual models. The basic idea behind this approach is that models that agree poorly with observations get less weight and models that largely duplicate existing models also get less weight. This is in line with the model averaging approaches discussed earlier, which also give models that agree poorly with observations less weight.

5 Conclusion

6 Acknowledgements

to be added after review

References

Artelle, K. A., S. C. Anderson, J. D. Reynolds, A. B. Cooper, P. C. Paquet, and C. T. Darimont (2016). Ecology of conflict: marine food supply affects human-wildlife interactions on land. *Scientific Reports 6*(1), 25936.

Baraldi, P., F. Mangili, and E. Zio (2013, April). Investigation of uncertainty treatment capability of model-based and data-driven prognostic methods using simulated data. *Reliability Engineering & System Safety 112*, 94–108.

Bishop, C. H. and G. Abramowitz (2013). Climate model dependence and the replicate Earth paradigm. *Climate Dynamics 41*(3), 885–900.

Bradley, S., R. Frigg, H. Du, and L. A. Smith (2014). Model Error and Ensemble Forecasting: A Cautionary Tale. *Scientific Explanation and Methodology of Science 1*, 58–66.

Brauer, F. (2008). Compartmental model in mathematical epidemiology. In F. Brauer, P. van den Driessche, and J. Wu (Eds.), *Mathematical Epidemiology*, Volume 1945, pp. 10–79. Springer.

Buckland, S. T., K. P. Burnham, and N. H. Augustin (1997, June). Model selection: An integral part of inference. *International Biometric Society 53*(2), 603–618.

Clyde, M. and E. I. George (2004). Model Uncertainty. *Statist. Sci. 19*(1), 81–94.

Cubasch, U., D. Wuebbles, D. Chen, M. C. Facchini, D. Frame, N. Mahowald, and J. G. Winther (2013). Introduction. In T. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, and P. Midgley (Eds.), *Climate*

Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, pp. 119–158. Cambridge; New York: Cambridge University Press.

Fernández, C., E. Ley, and M. F. J. Steel (2001). Model uncertainty in cross-country growth regressions. *Journal of Applied Econometrics 16*(5), 563–576.

Fragoso, T. M. and F. L. Neto (2015). Bayesian model averaging: A systematic review and conceptual classification. pp. 1–35.

Frigg, R., S. Bradley, H. Du, and L. A. Smith (2014). Laplace's Demon and the Adventures of His Apprentices. *Philosophy of Science 81*(1), 31–59.

Frigg, R., S. Bradley, R. L. Machete, and L. A. Smith (2013a). Probabilistic Forecasting: Why Model Imperfection Is a Poison Pill. In H. Andersen, D. Dieks, W. J. Gonzalez, T. Uebel, and G. Wheeler (Eds.), *New Challenges to Philosophy of Science*, pp. 479–491. Dordrecht: Springer Netherlands.

Frigg, R., L. A. Smith, and D. A. Stainforth (2013b). The Myopia of Imperfect Climate Models: The Case of UKCP09. *Philosophy of Science 80*(5), 886–897.

Frigg, R., L. A. Smith, and D. A. Stainforth (2015). An assessment of the foundational assumptions in high-resolution climate projections: the case of UKCP09. *Synthese 192*(12), 3979–4008.

Gelman, A. (2002). Prior distribution. *Encyclopedia of Environmetrics 3*, 1634–1637.

Goodwin, W. M. and E. Winsberg (2016). Missing the Forest for the Fish: How Much

Does the 'Hawkmoth Effect'Threaten the Viability of Climate Projections? *Philosophy of Science 83*(5), 1122–1132.

Hayashi, S. (1997). Connecting Invariant Manifolds and the Solution of the C1 Stability and Ω-Stability Conjectures for Flows. *Annals of Mathematics 145*(1), 81–137.

Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky (1999). Bayesian model averaging: A tutorial. *Statistical Science 14*(4), 382–401.

Jasny, B. R. and R. Stone (2017). Prediction and its limits. *Science 355*(6324), 468–469.

Jun, M., R. Knutti, and D. W. Nychka (2008, sep). Local eigenvalue analysis of CMIP3 climate model errors. *Tellus A 60*(5), 992–1000.

Knutti, R. (2010). The end of model democracy? *Climatic Change 102*(3), 395–404.

Knutti, R., R. Furrer, C. Tebaldi, J. Cermak, and G. A. Meehl (2010, dec). Challenges in Combining Projections from Multiple Climate Models. *Journal of Climate 23*(10), 2739–2758.

Knutti, R., J. Sedláček, B. M. Sanderson, R. Lorenz, E. M. Fischer, and V. Eyring (2017). A climate model projection weighting scheme accounting for performance and interdependence. *Geophysical Research Letters 44*(4), 1909–1918.

Ley, E. and M. F. J. Steel (2009). On the effect of prior assumptions in Bayesian model averaging with applications to growth regression. *Journal of Applied Econometrics 24*(4), 651–674.

Lorenz, E. N. (1972). Does the Flap of a Butterfly's wings in Brazil Set Off a Tornado in Texas? In *American Association for the Advancement of Science*, Cambridge MA.

Madigan, D. and A. E. Raftery (1994). Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window. *Journal of the American Statistical Association 89*(428), 1535–1546.

Mañé, R. (1987). A proof of the C^1 stability conjecture. *Publications Mathématiques de l'IHÉS 66*, 161–210.

Martin, M. A. and S. Roberts (2006). Bootstrap model averaging in time series studies of particulate matter air pollution and mortality. *Journal of Exposure Science and Environmental Epidemiology 16*, 242–250.

Montgomery, J. M. and B. Nyhan (2010, March). Bayesian Model Averaging: Theoretical Developments and Practical Applications. *Political Analysis 18*(2), 245–270.

Oreskes, N., D. A. Stainforth, and L. A. Smith (2010). Adaptation to Global Warming: Do Climate Models Tell Us What We Need to Know? *Philosophy of Science 77*(5), 1012–1028.

Palis, J. and S. Smale (1970). Structural stability theorems. *Proc. of Symposia in Pure Mathematic XIV*, 223–231.

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

Parker, W. S. (2011). When Climate Models Agree: The Significance of Robust Model Predictions. *Philosophy of Science 78*(4), 579–600.

Polley, E., E. LeDell, C. Kennedy, S. Lendle, and M. van der Laan (2017). *SuperLearner: Super Learner Prediction.* R package version 2.0-22.

Polley, E. C. and M. J. van der Laan (2010). Super learner in prediction. *U.C. Berkeley Division of Biostatistics Working Paper Series* (266).

Pope, V., S. Brown, R. Clark, M. Collins, W. Collins, C. Dearden, J. Gunson, G. Harris, C. Jones, A. Keen, J. Lowe, M. Ringer, C. Senior, S. Sitch, M. Webb, and S. Woodward (2007). The Met Office Hadley Centre climate modelling capability: the competing requirements for improved resolution, complexity and dealing with uncertainty. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 365*(1860), 2635–2657.

Puma, M. J., R. D. Koster, and B. I. Cook (2013). Phenological versus meteorological controls on land-atmosphere water and carbon fluxes. *Journal of Geophysical Research: Biogeosciences 118*(1), 14–29.

Randall, D. A., R. A. Wood, S. Bony, R. Colman, T. Fichefet, J. Fyve, V. Kattsov, A. Pitman, J. Shukla, J. Srinivasan, R. J. Stouffer, A. Sumi, and K. E. Taylor (2007). Climate Models and Their Evaluation. *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change.*

Roberts, S. and M. A. Martin (2010, January). Bootstrap-after-bootstrap model averaging for reducing model uncertainty in model selection for air pollution mortality studies. *Environmental Health Perspectives 118*(1), 131–136.

Saltelli, A., M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, and S. Tarantola (2008). *Global Sensitivity Analysis: The Primer.* John Wiley & Sons Ltd.

Seneviratne, S. I., M. G. Donat, A. J. Pitman, R. Knutti, and R. L. Wilby (2016). Allowable CO2 emissions based on regional and impact-related climate targets. *Nature 529*(7587), 477–483.

Sexton, D. M. H., J. M. Murphy, M. Collins, and M. J. Webb (2012). Multivariate probabilistic projections using imperfect climate models part I: outline of methodology. *Climate Dynamics 38*(11), 2513–2542.

Shankar, J., N. V. Solis, S. Mounaud, S. Szpakowski, H. Liu, L. Losada, W. C. Nierman, and S. G. Filler (2015). Using Bayesian modelling to investigate factors governing antibiotic-induced Candida albicans colonization of the GI tract. *Scientific Reports 5*, 8131.

Smale, S. (1966). Structurally Stable Systems are not Dense. *American Journal of Mathematics 88*(2), 491–496.

Smith, L. A. (2002, feb). What might we learn from climate forecasts? *Proceedings of the National Academy of Sciences 99*(suppl 1), 2487–2492.

Tebaldi, C. and R. Knutti (2007, aug). The use of the multi-model ensemble in probabilistic climate projections. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 365*(1857), 2053–2075.

Tebaldi, C., L. Mearns, D. Nychka, and R. Smith (2004, dec). Regional probabilities of precipitation change: A Bayesian analysis of multimodel simulations. *Geophysical Research Letters 31*(24).

van der Laan, M. J., E. C. Polley, and A. E. Hubbard (2007). Super learner. *Statistical Applications in Genetics and Molecular Biology 6*(1).

Wasserman, L. (2000). Bayesian Model Selection and Model Averaging. *Journal of Mathematical Psychology 44*, 92–107.

Weigel, A. P., R. Knutti, M. A. Liniger, and C. Appenzeller (2010, mar). Risks of Model Weighting in Multimodel Climate Projections. *Journal of Climate 23*(15), 4175–4191.