Katie Steele and Charlotte Werndl

The diversity of model tuning practices in climate science

Article (Published version)
(Refereed)

Original citation:
Steele, Katie and Werndl, Charlotte (2016) *The diversity of model tuning practices in climate science.* Philosophy of Science, 83 (5). pp. 1133-1144. ISSN 0031-8248

DOI: 10.1086/687944

The Diversity of Model Tuning Practices in Climate Science

Katie Steele and Charlotte Werndl*†

Many examples of calibration in climate science raise no alarms regarding model reliability. We examine one example and show that, in employing classical hypothesis testing, it involves calibrating a base model against data that are also used to confirm the model. This is counter to the 'intuitive position' (in favor of use novelty and against double counting). We argue, however, that aspects of the intuitive position are upheld by some methods, in particular, the general cross-validation method. How cross-validation relates to other prominent classical methods such as the Akaike information criterion and Bayesian information criterion is also discussed.

1. Introduction.

Many climate scientists are apprehensive about *calibrating* (or tuning) climate models to increase their reliability. This practice is commonly identified with including *parameterizations* in a climate model that 'stand in' for physical processes such as the behavior of clouds that are not well enough understood or are smaller than the grid size. The worry is that

*To contact the authors, please write to: Katie Steele, Australian National University; e-mail: katie.steele@anu.edu.au. Charlotte Werndl, University of Salzburg and London School of Economics; e-mail: charlotte.werndl@sbg.ac.at.

†We are grateful for valuable discussion and suggestions from the anonymous referees as well as to the audiences at the PSA 2014, the 2015 Philosophy of Climate Science Conference at the University of Pittsburgh, the Theory Construction in Science Conference at the London School of Economics, the Philosophy Colloquium at the University of Groningen, the Philosophy of Science Seminar at Bristol University, the Colloquium in Mathematical Philosophy at the Munich Center for Mathematical Philosophy, the British Society for the Philosophy of Science Seminar, the 2014 Trends in Logic Workshop at Ghent University, and the Third Reasoning Club Conference at the University of Kent. Funding support for the research was provided by the Arts and Humanities Research Council (AH/J006033/1) and by the ESRC Centre for Climate Change Economics and Policy, funded by the Economic and Social Research Council (ES/K006576/1 to Charlotte Werndl). Katie Steele was also supported by a 3-month research fellowship in residence at the Swedish Collegium for Advanced Study.

Philosophy of Science, 83 (December 2016) pp. 1133–1144. 0031-8248/2016/8305-0041$10.00

parameterizations are selected specifically to enhance the fit to the relevant observational data (say, change in global surface air temperature throughout the twentieth century), that is, to compensate in an ad hoc way for other structural errors in the model (see Frisch 2015 and references therein). Given this understanding of 'calibration', it is no wonder that climate scientists are skeptical about it. Indeed, the consensus, as echoed in the Intergovernmental Panel on Climate Change Fifth Assessment Report (AR5), is apparently that empirical fit with the calibrating data provides little to no confirmation for the calibrated model: "Model tuning directly influences the evaluation of climate models, as the quantities that are tuned cannot be used in model evaluation. Quantities closely related to those tuned will provide only weak tests of model performance. Nonetheless, by focusing on those quantities not generally involved in model tuning while discounting metrics clearly related to it, it is possible to gain insight into model performance" (Flato et al. 2013, box 9.1).

We dub this the 'intuitive position' regarding calibration and confirmation of base models/theories: that *use-novel* data have a special role in confirmation and, more strongly, that data cannot be used twice, both for calibration and confirmation (the *no-double-counting* rule; Worrall 2010).[1] We suggest, however, that scientists and philosophers alike overlook the diversity of model-calibration practices in science. Once one moves beyond highly suggestive examples, it is not obvious that the intuitive position is right. In the suggestive examples, calibration amounts to model construction that is ad hoc. Indeed, whether it is a calibrated version of Ptolemy's theory to fit planetary retrogressions or a calibrated climate model to fit the temperature record, the problem is that the adjustments to the base model have dubious prima facie reliability. If we think of calibration in this way—making ad hoc adjustments in order to get better fit—scientists are right to be skeptical about whether there is a net increase in reliability. The gain in empirical fit with the calibrating data must be traded off against the loss of physical plausibility of the model.

More 'modest' calibration examples in climate and other sciences provide better grounds for examining detailed questions of empirical fit vis-à-vis model confirmation. Indeed, there is much model calibration in climate science that is subtly at odds with the intuitive position: cases whereby data are used for both calibration and confirmation. Elsewhere, Steele and Werndl (2013) examined the calibration of the aerosol forcing parameter. In section 2 we will describe an even more pedestrian example of double counting in climate science. This sort of case differs from those above in that the 'tuned' base

[1] In this article we use the phrase 'confirmation' broadly, as pertaining to assessments of model reliability. Strictly speaking, confirmation (in philosophy) is about the truth of hypotheses.

models are not obviously inferior to any other base models under consideration. We can thus focus purely on the significance of model fit with calibration data.

The main focus of the article is the diversity of formal calibration methods and how these relate to the intuitive position. Steele and Werndl (2013) critique the intuitive position from the Bayesian perspective. Here we focus on the diverse class of classical or frequentist methods. Section 3 shows that the simplest classical method of hypothesis testing, employed in the case study described in section 2, is at odds with the intuitive position and is, in this regard, very similar to the Bayesian method. Section 4 introduces the general method of *cross-validation*, which allows for a more nuanced stance with respect to the intuitive position. Cross-validation can be refined depending on what frequentist properties of model assessment (*estimation* or *identification*) are considered desirable. In sections 4.1 and 4.2 we examine two methods, known as the Akaike and the Bayesian information criterion, which can each be related to special cases of cross-validation that accord with one or the other of these aims. The article concludes in section 5.

2. A Climate Case Study.

Stone et al. (2007) aim to explain global mean surface temperature changes in the past decades. A *climate forcing* measures the change in the net (downward minus upward) radiative flux at the top of the atmosphere or at the boundary between the troposphere and the stratosphere arising because of a change in an external driver of the climate system. Stone and colleagues look at four such forcings: (1) one associated with tropospheric greenhouse gases, (2) one associated with sulfate emissions, (3) one associated with stratospheric volcanic aerosols, and (4) one associated with solar radiation.

There is a specific spatiotemporal pattern of temperature changes (called a *fingerprint*) associated with each forcing. In Stone et al. (2007) the fingerprint for each forcing is known from energy balance models that incorporate information from a general circulation model. Yet what is not known is the relative extent of the response to a forcing. Hence, the extent of the response to a forcing corresponds to a free parameter that has to be estimated from the data.

Stone et al. (2007) use observations of mean surface temperature changes from 1940 to 2005 to estimate these free parameters and to measure the fit of the models with the observations. Estimating the relative extent of response to the various forcings amounts to finding the values of β_i, $1 \leq i \leq 4$, that give the best fit to the data T_{obs}. That is, the β_i in the following equation are fitted to minimize, within bounds, the error:

$$T_{\text{obs}} = \sum_{i=1}^{4} \beta_i T_i + \varepsilon, \tag{1}$$

where T_i is the pattern of temperature change given by the fingerprint for forcing i and ε is the error term.

More specifically, the base models Stone et al. (2007) consider include equation (1) and all other nested models derived from this (all in all 16 base models). That is, there is a base model in which there is only one free parameter β_1 corresponding to the extent of the response to the greenhouse gases forcing, a base model in which there are two free parameters corresponding to the extent of the response to the greenhouse gases forcing (β_1) and the sulfate emissions forcing (β_2), as well as base models with the other combinations of free β_i-parameters. *Model instances* are obtained when the free parameters are assigned specific values.

Stone et al. (2007) then compare the performance of these 16 base models, assuming that inclusion of the term $\beta_i T_i$ is necessary just in case the estimated β_i is significantly different from zero (at the 95% level). Stone and colleagues conclude that the base model $M_{1,2,3}$ that includes the three free parameters corresponding to the extent of the greenhouse gases forcing (β_1), the extent of the sulfate emissions forcing (β_2), and the extent of the stratospheric volcanic aerosols forcing (β_3) is confirmed relative to all other base models. Confidence intervals for the estimates of the greenhouse gases forcing, the sulfate emissions forcing, and the stratospheric volcanic aerosol forcing are provided. Stone and colleagues emphasise that this demonstrates that both anthropocentric and natural forcings are needed to account for the observations.

To sum up, Stone et al. (2007) use data about global mean temperature changes to estimate the values of the free parameters (*calibration*) and to confirm $M_{1,2,3}$ relative to the other 15 base models. That is, they engage in double counting, and use novelty does not play a role. The next section will reflect on the case study.

3. Classical Hypothesis Testing Vis-à-Vis the Intuitive Position. The simplest classical method for assessing models is arguably standard hypothesis testing—the procedure employed in our case study. We first describe classical hypothesis testing before turning to our case study.

A base model in this context is a set of model hypotheses that all share the same model structure (model equations) but that differ in the value of parameters that are considered the free parameters. These are thus referred to as *model-instance* hypotheses. Commonly, the dependent variable or model output is hypothesized to be an accurate representation of some aspect of the world. For climate models, the model output might represent, say, mean global temperature change. Hypothesis testing concerns one base model, although this includes any nested base models (i.e., subsets of the full set of model-instance hypotheses, where the value for one or more of the free parameters is zero).

Hypothesis testing considers whether the observational data are 'in keeping' with one or more of the model-instance hypotheses. If so, these hypotheses are treated as plausible candidates for the truth. What confers reliability is the testing procedure: as for all classical methods, long-run properties matter. More specifically, the hypothesis-testing procedure is as follows: all model-instance hypotheses for which the n data at hand fall in the unlikely or rejection region are discarded. The remaining accepted hypotheses form a confidence interval of plausible parameter values. The long-run properties of the testing procedure of interest are the type I error or significance level and the corresponding confidence level. It is assumed that the set of model-instance hypotheses under consideration form a suitable continuum, and the true hypothesis is among them. As such, the type I error is the (long-run frequentist) probability of rejecting any given model-instance hypothesis when it is in fact true (typically set at 0.05 or 0.01). The confidence level is the flip side of the type I error; the two values add to 1; the confidence level gives the (long-run frequentist) probability that the set of accepted model hypotheses, that is, the confidence interval for the various parameter values, contains the true hypothesis/parameter values, if the same testing procedure (with n data) were repeated indefinitely.

We return to our climate science example. Recall that the base model here is a linear combination of the 'fingerprints' of the various forcings (denoted by T_i):

$$T_{\text{obs}} = \sum_{i=1}^{4} \beta_i T_i + \varepsilon, \tag{2}$$

where the free parameters, β_i, $1 \leq i \leq 4$, indicate the extent of the forcings (and ε specifies probabilistic model error). The data are records of mean global temperature changes for the given time period, T_{obs}.

Hypothesis testing treats every possible combination of β_i values associated with the base model as a model-instance hypothesis. Any hypothesis for which the observed temperature record is too 'unlikely' (with type I error set at 0.05) is discarded, yielding a 95% confidence interval for the true model-instance hypothesis, which can be articulated in terms of 95% confidence intervals for each of the four β_i terms. It turned out that three of these β_i confidence intervals did not contain zero: the β_i associated with the greenhouse gases forcing, the sulfate emissions forcing, and the stratospheric volcanic aerosol forcing. Thus, the base model that includes these parameters is deemed more reliable (or confirmed) relative to the nested base models that do not include these parameters (effectively setting them to zero).[2]

[2] This is a weaker conclusion than those of Stone et al. (2007). By our analysis, the base models that do not have positive values for β_1, β_2, and β_3 are falsified; all other base models are consistent with the data.

Contrary to the intuitive position, classical hypothesis testing does not respect use novelty and the no-double-counting rule: calibration is the assessment of particular model-instance hypotheses—these hypotheses are either accepted in the confidence interval or rejected. When forming a confidence interval, one base model may be accepted over another otherwise nested base model (when the confidence interval for some free parameter does not include zero).[3] Thus, there is double counting, and data used for confirmation are not use novel.

4. Cross-Validation Vis-à-Vis the Intuitive Position. While hypothesis testing may be the most widely used classical method, other classical methods have been proposed. For all these methods, the focus is the long-run properties of the procedure that is used to assess/identify models. The hypothesis tester asks herself: What confidence level, $1 - \alpha$, is suitable for my purposes, given that if I were to repeat this procedure indefinitely, then my confidence interval would contain the true hypothesis in $(1 - \alpha)\%$ of cases? The crucial assumption is that the base model (or otherwise a nested counterpart) is true. But there may be contexts in which the scientist is not sure which base model is true, and the plausible candidates do not simply amount to a nested family of base models. In this case, hypothesis testing is not very telling—we are assured only of the long-run accuracy of the confidence interval for each base model, conditional on that base model being true. This does not license any comparison of base models, unless they are nested. The cross-validation method, by contrast, is more general. It also sheds a different light on the intuitive position with respect to use novelty and double counting.

Cross-validation is a general method for assessing/identifying models for prediction, which has also been applied and discussed in climate science (e.g., Michaelsen 1987; Elsner and Schwertmann 1994). It has several main components that can be adjusted, depending on the context and the desired long-run properties. The first component is the procedure that is being assessed for each base model. This is the calibration step and is akin to the hypothesis-testing procedure but is generally an abbreviated version whereby what is identified is just the model instance for each base model that gives the best fit with (confers highest probability to) the n data points. This is referred to as the *maximum likelihood* estimator for the base model. The second component is the performance measure for the base-model procedure. Typically it is the mean predictive accuracy (with respect to predicting a new data point) of the base-model procedure, if it were conducted indef-

[3] Admittedly, the most inclusive base model is simply assumed true in hypothesis testing and so cannot be confirmed or disconfirmed. Subsets of this base model may, however, be confirmed relative to other subsets.

initely in response to n data generated randomly by nature. Given that we do not know nature's data-generating mechanism, we must estimate the mean predictive accuracy of the base-model procedure. The way this estimate is determined is the key characteristic of any model selection method.

The typical $(n - 1)$-cross-validation estimator for the mean predictive accuracy is calculated as follows: Given n data points, one starts by using the first $n - 1$ data points to construct the best-fitting model instance of the base model given these data and then uses the remaining data points to assess the performance of the model instance (by calculating the distance between the predicted data point and the actual data point). This is repeated for all possible selections of $n - 1$ data points to calculate the mean distance between the predicted and actual data points. An alternative is $(n - k)$-cross-validation, where $n - k$ data points are used to find the best-fitting model instance, and the remaining k data points are used to assess predictive accuracy. The key assumption is that the data are independently and identically distributed (Arlot and Celisse 2010).

Unlike hypothesis testing, cross-validation gives use-novel data a special standing. It effectively involves repeated tests whereby one or more data points are 'left out of calibration' to serve as the telling novel data. However, cross-validation does not respect the no-double-counting rule: all data are used for confirmation and calibration.

The cross-validation estimators of the long-run predictive accuracy of base-model procedures themselves have long-run properties. One property is the bias: how well the expected estimate of predictive accuracy matches the true predictive accuracy of the maximum-likelihood procedure. The smaller the value for k in $(n - k)$-cross-validation, the less biased the estimator. The $(n - 1)$-cross-validation estimator, for instance, is an asymptotically unbiased estimator (Linhart and Zucchini 1986; Zucchini 2000; Arlot and Celisse 2010). For larger values of k, we get biased estimates because we are assessing the performance of the base-model procedure when $n - k$ data points are used for calibration and not what one would like to test: the performance of the procedure when n data points are used for calibration (as is actually done).

Whether one should opt for a biased or an unbiased estimator of predictive accuracy is related to the question of one's aims in model selection (Arlot and Celisse 2010). A method is *efficient* if, as the number of data points, n, approaches infinity, the probability approaches one that the base-model procedure (maximum likelihood estimator) with greatest predictive accuracy is selected. This property characterizes the goal of estimation. A method is *model consistent* if, as the number of data points, n, approaches infinity, the probability approaches one that the true model instance is selected. This property characterizes the goal of identification. As it happens, the usual situation is that it is not possible for a cross-validation method to have both properties. Indeed, efficiency corresponds to an unbiased estimator, while model consistency corresponds to a biased estimator. In what follows, we analyze this distinction further by relating cross-validation to two well-known model selection methods: the Akaike information criterion (AICc) and the Bayesian information criterion (BIC). This comparison also allows a clearer picture of how AICc and BIC measure up with respect to use novelty and no double counting.

4.1. Comparison to the Akaike Information Criterion. The AICc for finite sample sizes aims at estimation, that is, to determine the base-model procedure that performs best for predictive tasks. For AICc the distance between the actual and the simulated observations is measured by the Kullback-Leibler discrepancy,[4] and the data have to be independently and identically distributed (there are some further technical assumptions; see Linhart and Zucchini 1986; Burnham and Anderson 1998).

As usual in model selection, AICc estimates the predictive accuracy of the maximum likelihood estimator. So the calibration step is to identify the best-fitting model instance for each base model relative to the n data points; these are the model instances that would be used for prediction. To estimate the long-run average predictive accuracy of each base-model procedure, first the discrepancy between the best-fitting model instance and the actual data points is calculated. It is $-\ln[L]/n$, where L is the maximum of the likelihood function (Zucchini 2000, 52–53). The following expression then gives the score estimating the average predictive accuracy of the base-model procedure given n data points:

$$C_{\mathrm{AICc}} = -\frac{\ln[L]}{n} + \left(\frac{p}{n} + \frac{p(p+1)}{n(n-p-1)}\right), \qquad (3)$$

where p is the number of free parameters. C_{AICc} can be shown to be an unbiased estimator (Linhart and Zucchini 1986; Burnham and Anderson 1998).

Clearly, for AICc there is double counting: all the data are used first for calibration and then for confirmation (i.e., to calculate the score [3]). Also, clearly, the data used for confirmation are not use novel since the maximum likelihood given all the data is used for calculating the score (3). So, in contrast to cross-validation, there is no apparent assessment of how the base-model procedure fares on new data. Despite this, in a precise sense, there is a penalty term in the expression for the estimation of the predictive accuracy (3) because of the data already having been used for calibration. To demonstrate this, we now compare two methods for estimating the average predictive accuracy of procedures where n data points are used for calibra-

[4] Our conceptual points carry over to other distance measures.

tion with the only difference that (A) in the first case the data used for confirming the procedure are use novel, and (B) in the second case they are not.

We start with case A. Here one first engages in calibration; that is, one uses the n data to determine the model instance that fits the data best. Then with n novel data points the distance between the predicted and the actual data points is calculated in order to estimate the average predictive accuracy of the procedure (n data points are considered because later we compare this method to the AICc, where also n data points are used for confirmation). As explained above, this yields an unbiased estimator of the average predictive accuracy of the maximum likelihood estimator constructed from n data points (Linhart and Zucchini 1986; Zucchini 2000).[5]

We now turn to AICc and case B, where the data are not use novel. One starts as in A and uses n data points for calibration to determine the best model instance. Yet for confirmation one now uses the same n data points that have been used for calibration before (hence, these are not use novel). More specifically, these n data points are used exactly as in A to determine the average Kullback-Leibler divergence between the n data points and the best-fitting model instance. In this way one obtains the term on the left-hand side of C_{AICc}. Note that the way we proceeded so far has been exactly as in A, with the only difference that the data are not use novel. Yet the term on the left-hand side of C_{AICc} would lead to a statistically very biased estimate (the fit is assessed by the same data that have been used to determine the model instance and is thus likely to be better than if novel data had been used). In order to obtain an unbiased estimator (when n data points are used for calibration), the term on the right-hand side of C_{AICc} is needed. Consequently, the term on the right-hand side amounts to a penalty term because the data have already been used for calibration.

In sum: for AICc the data are not use novel, and there is double counting. Still, comparison with cross-validation yields that use novelty plays a role: since the data have already been used for calibration, there is a penalty term in the score that measures confirmation.[6]

4.2. Comparison to the Bayesian Information Criterion. The BIC, in contrast to AICc, aims at identification of the true model. Indeed, as the name suggests, BIC is purportedly a Bayesian approach that aims to assess base models in terms of their comparative posterior probabilities. The posteriors for base models are measured in terms of the marginal likelihoods of the base-model hypotheses (the weighted average of the likelihoods for the

[5] The estimator would also be unbiased if more than n data points were used for confirmation.

[6] Another important result is that $(n - 1)$-cross-validation is asymptotically equivalent to the AICc (Stone 1977).

relevant model-instance hypotheses). The marginal likelihoods track the posterior probabilities just in case the prior probabilities for the base models are the same (or in case n is very large, such that the prior probabilities have negligible importance).

The BIC score for a base model is eventually an approximation of -2 times the log of the marginal likelihood of the base model. It is assumed that the likelihood probability density functions (with regard to the Lebesgue measure μ) belong to the exponential family. The approximating expression is as follows (for derivation, see Schwarz 1978; reproduced in Sprenger 2013):

$$\text{BIC} = -2 \times \ln[L] + k \ln[n], \tag{4}$$

where L is the maximum of the likelihood function (the likelihood for the maximum likelihood estimator), n is the number of data, and k is the number of free parameters for the base model. In short, the term $k \ln[n]$ corrects for the fact that the likelihood for the maximum likelihood estimator overestimates the marginal likelihood for the base model, in a way dependent on both the number of free parameters and the number of data. The lower the BIC score, the more ‘choice-worthy’ the base model.

Strictly speaking, BIC assesses models in terms of their marginal likelihoods rather than their posterior probabilities; it is only in special cases that the two yield the same results (cf. Sober 2008; Romeijn, van der Schoot, and Hoijtink 2012). Indeed, where nested models are concerned, the base models do not have the same prior probabilities (except for trivial cases), and neither prior nor posterior probabilities will ever favor the more nested base model, since logic dictates that it has lesser probability than any wider base model that it entails. Thus, BIC is not exactly Bayesian, because it is unclear why a Bayesian should care about the relative marginal likelihoods of base models if these do not track posterior probabilities.[7] Indeed, the use of BIC to compare models is generally justified in terms of the frequentist properties of this method, such as model consistency, as discussed above—hence, our grouping of BIC with classical model selection methods.

As for AICc, there is double counting for BIC because all the data are used for calibration and confirmation (i.e., to determine the score [4]), and the data used for confirmation are not use novel since the maximum of the likelihood function is used to calculate the score (4). Still, as for AICc, one can compare BIC with cross-validation to see that $k \times \ln[n]$—the term on the right-hand side of equation (4)—corresponds to a penalty term due to the data having already been used for calibration (although the comparison

[7] It is also not clear why Bayesians should care about sets of model hypotheses rather than individual model hypotheses.

is much less general because it has been established rigorously only for certain cases, including linear regression; Arlot and Celisse 2010). More specifically, for linear regression, $(n - k)$-cross-validation (where use novelty is important), when k/n goes to 1 as n goes to infinity, is consistent (Arlot and Celisse 2010). BIC is consistent too, and by comparing it to $(n - k)$-cross-validation when k/n goes to 1 as n goes to infinity, we see that the term $k \times \ln[n]$ can be interpreted as a penalty term because the data have already been used before for calibration.[8]

In sum: for BIC the data are not use novel, and there is double counting. Still, for certain cases there is at least a role for use novelty in the sense that there is a penalty term in the score that measures confirmation, because of the data having been used already for calibration.

5. Conclusion. This article focused on the diversity of calibration methods and how these relate to the 'intuitive position', which claims that data for confirmation have to be use novel and that double counting (using the same data for calibration and confirmation) is illegitimate. We first showed that the simplest classical method of hypothesis testing (employed in many climate science papers) is at odds with the intuitive position. Then we discussed the general method of cross-validation, which presented us with a more nuanced stance with respect to use novelty and double counting: here use novelty is important, but there is still double counting. Cross-validation can be refined depending on what frequentist properties of model assessment (estimation or identification) are considered desirable. Finally, we compared cross-validation with the AICc and BIC: in this way we have seen that while for these criteria the data are not use novel, the idea of novel data is still relevant in the sense that there is a penalty term in the score that measures confirmation because the data have been used for calibration before.

Our discussion has normative bearing in the following sense: if the intuitive position is inconsistent with prominent formal methods of calibration, as we have shown here, so much the worse for the intuitive position. At the very least, this suggests that the intuitive position must be refined. We leave the question open as to whether the most minimal refinement of the intuitive position (arguably the class of cross-validation methods) is to be preferred, normatively speaking.

[8] For AICc and cross-validation, the comparison is neater (it is more general), and for both cross-validation and AICc, the number of data points (n) used for calibration and for confirmation is the same. In contrast, when comparing $(n - k)$-cross-validation with BIC, the number of data points is different: for cross-validation, $n - k$ data are used for calibration and then k for confirmation, but for BIC, n data are used for calibration and confirmation.

REFERENCES

Arlot, Sylvain, and Alain Celisse. 2010. “A Survey of Cross-Validation Procedures for Model Selection.” *Statistics Surveys* 4:40–79.

Burnham, Kenneth P., and David R. Anderson 1998. *Model Selection and Multimodal Inference*. Berlin: Springer.

Elsner, James B., and Carl P. Schwertmann. 1994. “Assessing Forecast Skill through Cross-Validation.” *Weather and Forecasting* 9:619–24.

Flato, G., J. Marotzke, B. Abiodun, P. Braconnot, S. C. Chou, W. Collins, P. Cox, F. Driouech, S. Emori, V. Eyring, C. Forest, P. Gleckler, E. Guilyardi, C. Jakob, V. Kattsov, C. Reason, and M. Rummukainen. 2013. “Evaluation of Climate Models.” In *Climate Change, 2013: The Physical Science Basis; Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, ed. T. F. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S. K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, and P. M. Midgley. Cambridge: Cambridge University Press.

Frisch, Mathias. 2015. “Predictivism and Old Evidence: A Critical Look at Climate Model Tuning.” *European Journal for Philosophy of Science* 5 (2): 171–90.

Linhart, H., and Walter Zucchini. 1986. *Model Selection*. Wiley Series in Probability and Statistics. New York: Wiley.

Michaelsen, Joel. 1987. “Cross-Validation in Statistical Climate Forecast Models.” *Journal of Climate and Applied Meteorology* 26:1589–1600.

Romeijn, Jan-Willem, Rens van der Schoot, and Herbert Hoijtink. 2012. “One Size Does Not Fit All: Derivation of a Prior-Adapted BIC.” In *Probabilities, Laws, and Structures*, ed. Dennis Dieks, Wenceslao Gonzales, Stephan Hartmann, Fritz Stadler, Thomas Uebel, and Marcel Weber, 87–106. Berlin: Springer.

Schwarz, Gideon. 1978. “Estimating the Dimension of a Model.” *Annals of Statistics* 6:461–64.

Sober, Elliott. 2008. *Evidence and Evolution*. Cambridge: Cambridge University Press.

Sprenger, Jan. 2013. “The Role of Bayesian Philosophy within Bayesian Model Selection.” *European Journal for the Philosophy of Science* 3:101–14.

Steele, Katie, and Charlotte Werndl. 2013. “Climate Models, Confirmation and Calibration.” *British Journal for the Philosophy of Science* 64:609–35.

Stone, Daithi A., Myles R. Allen, Frank Selten, Michael Kliphuis, and Peter A. Stott. 2007. “The Detection and Attribution of Climate Change Using an Ensemble of Opportunity.” *Journal of Climate* 20:504–16.

Stone, M. 1977. “An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike’s Criterion.” *Journal of the Royal Statistical Society* B 39 (1): 44–47.

Worrall, John. 2010. “Error, Tests, and Theory Confirmation.” In *Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability, and the Objectivity and Rationality of Science*, ed. Deborah G. Mayo and Aris Spanos, 125–54. Cambridge: Cambridge University Press.

Zucchini, Walter. 2000. “An Introduction to Model Selection.” *Journal of Mathematical Psychology* 44:41–61.