Model confirmation in climate economics

Antony Millnera,1 and Thomas K. J. McDermottb,b

aGrantham Research Institute on Climate Change and the Environment, London School of Economics and Political Science, London WC2A 2AE, United Kingdom; and bSchool of Economics, University College Cork, Cork T12 YN60, Ireland

Edited by M. Granger Morgan, Carnegie Mellon University, Pittsburgh, PA, and approved June 7, 2016 (received for review March 19, 2016)

Benefit-cost integrated assessment models (BC-IAMs) inform climate policy debates by quantifying the trade-offs between alternative greenhouse gas abatement options. They achieve this by coupling simplified models of the climate system to models of the global economy and the costs and benefits of climate policy. Although these models have provided valuable qualitative insights into the sensitivity of policy trade-offs to different ethical and empirical assumptions, they are increasingly being used to inform the selection of policies in the real world. To the extent that BC-IAMs are used as inputs to policy selection, our confidence in their quantitative outputs must depend on the empirical validity of their modeling assumptions. We have a degree of confidence in climate models both because they have been tested on historical data in hindcasting experiments and because the physical principles they are based on have been empirically confirmed in closely related applications. By contrast, the economic components of BC-IAMs often rely on untestable scenarios, or on structural models that are comparatively untested on relevant time scales. Where possible, an approach to model confirmation similar to that used in climate science could help to build confidence in the economic components of BC-IAMs, or focus attention on which components might need refinement for policy applications. We illustrate the potential benefits of model confirmation exercises by performing a long-run hindcasting experiment with one of the leading BC-IAMs. We show that its model of long-run economic growth—one of its most important economic components—had questionable predictive power over the 20th century.

We focus on benefit-cost IAMs (BC-IAMs) in this article, because these have been the focus of research activity in economics (5, 6) and are increasingly influential in policy applications.

BC-IAMs couple simplified climate models with representations of the global economy in an attempt to understand the trade-offs between alternative policy options. They have been applied to a wide variety of questions. How might different welfare frameworks affect the attractiveness of policy options (7)? Which approaches to international agreements are likely to succeed (8)? How might different policy instruments affect innovation in energy technologies (9)? These modeling exercises provide valuable insights into the possible qualitative differences between policy options. However, their quantitative implications are conditional on the veracity of the underlying models. BC-IAMs can be used to show that policy A leads to higher welfare than policy B in model X, but to extend this model-based finding to claims about reality we need to know how well X approximates reality. Are the equations and initialization procedures used by model X structurally sound, and, if not, what risks might we run by treating them as if they are?

Assessing the structural soundness of economic modeling assumptions in BC-IAMs has recently become an increasingly pressing issue, because they are beginning to be used to inform quantitative real-world policy decisions. For example, the US government has recently established an interagency working group (10, 11) to estimate a value of the social cost of carbon (SCC), the welfare cost to society from emitting a ton of CO2. The value of the SCC that was adopted will form part of the cost-benefit assessment of all federal projects and policies, and thus has the potential to influence billions of dollars of investment. The process used to establish a value for the SCC relied heavily on BC-IAMs, the first time they have directly

Supporting Information

SUSTAINABILITY SCIENCE

Author contributions: A.M. and T.K.J.M. designed research; A.M. and T.K.J.M. performed research; A.M. and T.K.J.M. analyzed data; and A.M. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

*aGrantham Research Institute on Climate Change and the Environment, London School of Economics and Political Science, London WC2A 2AE, United Kingdom; and bSchool of Economics, University College Cork, Cork T12 YN60, Ireland

Supporting Information online at www.pnas.orglookup/10.1073/pnas.1604121113

Significance

Benefit-cost integrated assessment models (BC-IAMs) combine climate science, impacts studies, and representations of long-run economic growth to estimate the costs and benefits of climate policy. They provide valuable qualitative insights into how policy outcomes might depend on ethical and empirical assumptions. Increasingly, however, BC-IAMs are being used to inform quantitative policy choices, despite their economic components being largely untested, or untestable, over relevant time scales. We demonstrate the potential benefits of model confirmation exercises for policy applications, demonstrating that the economic growth model used by a prominent BC-IAM had little predictive power over the 20th century. Insofar as possible, out-of-sample empirical tests of the economic components of BC-IAMs should inform their future development for real-world applications.

Author contributions: A.M. and T.K.J.M. designed research; A.M. and T.K.J.M. performed research; A.M. and T.K.J.M. analyzed data; and A.M. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

*aGrantham Research Institute on Climate Change and the Environment, London School of Economics and Political Science, London WC2A 2AE, United Kingdom; and bSchool of Economics, University College Cork, Cork T12 YN60, Ireland

Supporting Information online at www.pnas.orglookup/10.1073/pnas.1604121113

Significance

Benefit-cost integrated assessment models (BC-IAMs) combine climate science, impacts studies, and representations of long-run economic growth to estimate the costs and benefits of climate policy. They provide valuable qualitative insights into how policy outcomes might depend on ethical and empirical assumptions. Increasingly, however, BC-IAMs are being used to inform quantitative policy choices, despite their economic components being largely untested, or untestable, over relevant time scales. We demonstrate the potential benefits of model confirmation exercises for policy applications, demonstrating that the economic growth model used by a prominent BC-IAM had little predictive power over the 20th century. Insofar as possible, out-of-sample empirical tests of the economic components of BC-IAMs should inform their future development for real-world applications.

Author contributions: A.M. and T.K.J.M. designed research; A.M. and T.K.J.M. performed research; A.M. and T.K.J.M. analyzed data; and A.M. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

*aGrantham Research Institute on Climate Change and the Environment, London School of Economics and Political Science, London WC2A 2AE, United Kingdom; and bSchool of Economics, University College Cork, Cork T12 YN60, Ireland

Supporting Information online at www.pnas.orglookup/10.1073/pnas.1604121113

Significance

Benefit-cost integrated assessment models (BC-IAMs) combine climate science, impacts studies, and representations of long-run economic growth to estimate the costs and benefits of climate policy. They provide valuable qualitative insights into how policy outcomes might depend on ethical and empirical assumptions. Increasingly, however, BC-IAMs are being used to inform quantitative policy choices, despite their economic components being largely untested, or untestable, over relevant time scales. We demonstrate the potential benefits of model confirmation exercises for policy applications, demonstrating that the economic growth model used by a prominent BC-IAM had little predictive power over the 20th century. Insofar as possible, out-of-sample empirical tests of the economic components of BC-IAMs should inform their future development for real-world applications.

Author contributions: A.M. and T.K.J.M. designed research; A.M. and T.K.J.M. performed research; A.M. and T.K.J.M. analyzed data; and A.M. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

*aGrantham Research Institute on Climate Change and the Environment, London School of Economics and Political Science, London WC2A 2AE, United Kingdom; and bSchool of Economics, University College Cork, Cork T12 YN60, Ireland

Supporting Information online at www.pnas.orglookup/10.1073/pnas.1604121113

Significance

Benefit-cost integrated assessment models (BC-IAMs) combine climate science, impacts studies, and representations of long-run economic growth to estimate the costs and benefits of climate policy. They provide valuable qualitative insights into how policy outcomes might depend on ethical and empirical assumptions. Increasingly, however, BC-IAMs are being used to inform quantitative policy choices, despite their economic components being largely untested, or untestable, over relevant time scales. We demonstrate the potential benefits of model confirmation exercises for policy applications, demonstrating that the economic growth model used by a prominent BC-IAM had little predictive power over the 20th century. Insofar as possible, out-of-sample empirical tests of the economic components of BC-IAMs should inform their future development for real-world applications.

Author contributions: A.M. and T.K.J.M. designed research; A.M. and T.K.J.M. performed research; A.M. and T.K.J.M. analyzed data; and A.M. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

*aGrantham Research Institute on Climate Change and the Environment, London School of Economics and Political Science, London WC2A 2AE, United Kingdom; and bSchool of Economics, University College Cork, Cork T12 YN60, Ireland

Supporting Information online at www.pnas.orglookup/10.1073/pnas.1604121113

Significance

Benefit-cost integrated assessment models (BC-IAMs) combine climate science, impacts studies, and representations of long-run economic growth to estimate the costs and benefits of climate policy. They provide valuable qualitative insights into how policy outcomes might depend on ethical and empirical assumptions. Increasingly, however, BC-IAMs are being used to inform quantitative policy choices, despite their economic components being largely untested, or untestable, over relevant time scales. We demonstrate the potential benefits of model confirmation exercises for policy applications, demonstrating that the economic growth model used by a prominent BC-IAM had little predictive power over the 20th century. Insofar as possible, out-of-sample empirical tests of the economic components of BC-IAMs should inform their future development for real-world applications.

Author contributions: A.M. and T.K.J.M. designed research; A.M. and T.K.J.M. performed research; A.M. and T.K.J.M. analyzed data; and A.M. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.
informed quantitative federal rules. Whereas the US SCC estimate is perhaps the most prominent recent example, other governments and international organizations are also increasingly turning to BC-IAMs to inform policy choices. As soon as a model is used to inform quantitative policy decisions, the criteria by which it must be judged become more demanding. A given model may be a useful tool for exploring the qualitative implications of different assumptions, but in order for it to be profitably applied to policy choices we need to know how plausible those assumptions are as empirical hypotheses about how the world works. If a model can be shown to be structurally flawed in hindcasting exercises, our expectation should be that similar errors might occur when using it to make predictions that inform policy choices today. No model is perfect, and we should not expect any given model to ensure us against regret entirely. However, to the extent possible, it is in our interests to attempt to ascertain how wrong we might go when relying on a model to make decisions. As has occurred in climate science, this exercise could build confidence in those economic modeling assumptions that are found to be consistent with empirical data, and focus attention on those assumptions that may require refinement for policy applications.

Importantly, confirmation exercises provide entirely different information about a model’s validity from model calibration, sensitivity analysis, probabilistic approaches to quantifying parametric uncertainty, or expert elicitation of model parameters, all of which are standard practice in the field. These uncertainty quantification methods explore the space of model outcomes (and perhaps estimate their likelihood), taking the model’s structural assumptions as given. Model confirmation, however, tests whether the equations and initialization procedures a model uses to generate predictions are able to provide a good representation of observed outcomes. A model whose outcome space has been explored using the uncertainty quantification methods mentioned above may still yield error-prone predictions if the underlying modeling assumptions are not a good fit to reality. Although these methods can of course generate a distribution of model outcomes, whether or not such distributions reflect the uncertainty we actually face depends on the structural soundness of the model used to generate them. We note that model confirmation is only possible if a model is specified in a self-contained manner (i.e., it is composed of a set of structural assumptions and free parameters that can, at least in principle, be estimated from data). This makes models that rely on fixed external scenarios for generating predictions very difficult to confirm ext ante. Although such models could provide a good characterization of current uncertainty, we have no way of assessing whether this is likely to be the case by testing their past performance.

Although the physical science models upon which the scientific components of BC-IAMs are based have often been subjected to tests of structural validity, their economic components are often either based on untestable exogenous scenarios, or on structural modeling assumptions that are largely untested on the temporal scales that are relevant to climate applications. In part this reflects genuine data difficulties, which make some economic assumptions in BC-IAMs very difficult to confirm. For example, BC-IAMs assume a functional form for the climate damage function, which quantifies the impact of global average temperature changes on the aggregate productivity of the economy. BC-IAM results are highly sensitive to the rate of increase of damages with temperature at high temperatures, but because we have only seen a small amount of average warming so far, it is very difficult to test any assumed functional form for damages. Some of the most important economic components of some BC-IAMs are, however, amenable to empirical tests. An Out-of-Sample Test of a Model of Long-Run Economic Growth

To demonstrate what may be learned from model confirmation exercises we focus on the economic growth model used by the well-known Dynamic Integrated Climate-Economy (DICE) BC-IAM (15). The assumptions BC-IAMs make about long-run economic growth have a very substantial effect on leading policy outputs such as the SCC. This is because economic growth strongly affects the path of greenhouse gas emissions, the magnitude of climate damages, and the wealth of future generations, all key determinants of the aggregate costs and benefits of climate policy. Unlike other well-known BC-IAMs (e.g., refs. 16 and 17), which rely on external scenarios for economic growth that are impossible to test empirically ex ante, DICE uses an explicit model of economic growth that makes it well-suited to empirical testing, and is also widely deployed across climate economics (see, e.g., ref. 18). A crucial part of this growth model is a model of the temporal evolution of total factor productivity (TFP), a quantity that sets the overall level of technological advancement in the economy. Economic growth is largely driven by technological progress in DICE. Thus, although policy evaluation in DICE is also highly sensitive to other structural modeling assumptions (e.g., the shape of the damage function and the evolution of abatement costs), a lot depends on how it models overall technological progress.

To test the structural assumptions and initialization procedures used by DICE’s economic growth model, we consider the following question: How would this model fare if we asked it to predict the growth path of a major economy over the 20th century? This question is closely analogous to those asked of climate models by climate scientists (2). The model of the evolution of the economy DICE employs is a version of the Ramsey neoclassical growth model, familiar to any student of macroeconomics (see ref. 19 for a detailed exposition). In this model economic output is generated by competitive firms and is either consumed or reinvested in firms. Firms produce output via a production technology, which uses the capital and labor supplied by consumers as inputs. In DICE technological progress is modeled as an increasing trend in TFP, which acts as a multiplier on firms’ production technologies. Thus, as TFP grows, and the technologies of production become more advanced, fewer capital and labor inputs are required to generate a given level of economic output. A specific model of the time dependence of TFP is assumed in DICE. This model depends on free parameters that can be estimated from economic history.

We test this model’s predictive performance using recently compiled data on the US economy from 1870 to 2010 (21). We single out the United States because it is the largest economy for which detailed long-run economic data are available, and because of its position at the technological frontier over much of

---

1 This view is confirmed by Nordhaus (12): “The major factor producing different climate outcomes in our uncertainty runs is differential technological change. In our estimates, the productivity uncertainty outweighs the uncertainties of the climate system and the damage function in determining the relationship between temperature change and consumption.” A general sensitivity analysis of the DICE model confirms that its SCC estimates are highly sensitive to the growth rate of TFP (13). A heuristic understanding of why policy recommendations are so sensitive to assumptions about TFP growth can be obtained by studying the social discount rate $\rho(t)$. Under standard assumptions the change in social welfare that arises from a small change in consumption $\Delta c_t$ that occurs $t$ years in the future is given by $\Delta w_t \approx \Delta c_t e^{-\rho(t)t}$. Standard computations (14) show that in a deterministic setting $\rho(t) = \delta + \gamma(t)$, where $\delta$ is the pure rate of social time preference, $\gamma$ is the elasticity of marginal utility, and $\gamma(t)$ is the average consumption growth rate between the present and time $t$. In most cases the term $\gamma(t)$ is the dominant contribution to $\rho(t)$. Because consumption growth $\gamma(t)$ is driven by TFP growth in DICE, the present value of future climate damages is highly sensitive to TFP growth. For example, for $\delta = 1\%$, $\gamma = 2$ and $g(100) = 1\%$, an incremental climate damage of $100$ that occurs in $100$ years from now will be valued at $-$0.01/0.01 + $1/0.01 - (1/0.01 + 1/0.01) = $95 in present value terms. For $g(100) \approx 2\%$ however, the same $100$ damage would be worth $100(1 - 0.02 + 0.02/2) = 90.5$. Thus, an increase in consumption growth from $1\%$ to $2\%$ reduces the current welfare cost of climate damages that occur in $100$ by a factor larger than $7$.

2 DICE assumes that the growth rate of TFP is an exponential function that slowly decays from an initial value to a smaller long-run value. The free parameters are the initial growth rate and the rate of decline of the growth rate. A large literature has developed endogenous growth models that relate the evolution of TFP to endogenous economic variables (see refs. 19 and 20 for a review of applications to climate economics). Because we wish to stay as close as possible to the methodology used by DICE, we do not investigate the empirical performance of these models here, but see comments below and Supporting Information, Endogenous Growth Versions of the DICE Model.
the 20th century (22). Our tests are as generous as possible to the model (e.g., we assume a perfect forecast of labor supply) and stick closely to the calibration and forecasting methodology used by DICE (details of the model implementation are available in Supporting Information). To test the model’s predictive power we divide the data into different training and verification windows; 95% confidence intervals (CI) for the parameters of the TFP model are inferred from the training data. The state equation for the capital stock and empirically estimated model of TFP evolution are then used to predict economic output.

Fig. 1 is illustrative of our model estimation and confirmation methodology. The figure depicts a long-run forecast of TFP and economic output obtained by estimating the TFP model on the 50 y of data from 1870 to 1920. Fig. 1, Left depicts the fit of the TFP model to the training data and its out-of-sample projection of TFP. Fig. 1, Right shows an out-of-sample projection of gross domestic product (GDP) at 1920, which is generated using the empirically estimated TFP model, the state equation for the evolution of the capital stock, and a perfect forecast of labor supply. The figure shows that although the TFP model fits the training data well, its out-of-sample forecast substantially underestimates technological progress in the latter half of the 20th century. These errors are compounded for GDP predictions, because persistent underestimates of TFP affect predicted investment flows and capital formation in each future period, which further bias the model downward. We note, however, that although the presence of model errors is significant, the fact that the model was downward-biased in 1920 does not imply that it will be downward-biased in all periods, as we demonstrate below.

Although the out-of-sample forecasts of TFP and GDP the model generates in 1920 are not successful, they nevertheless look reasonable when viewed from the perspective of the 50-y data series up to that year. Had economists produced these predictions at the time based on only these 50 y of data, they would no doubt have been perceived as plausible future growth scenarios. From today’s perspective, however, the model looks like a less-reliable predictive tool. An important reason why the model performs poorly is that the post-World War Two boost in productivity growth is not presaged in the training data at 1920. This finding is indicative of the difficulty of predicting long-run technological developments. We face precisely the same difficulties today when using BC-IAMs to project economic growth into the next century and beyond (see Supporting Information for further discussion).

Whereas Fig. 1 suggests that the model’s long-run predictive performance could be a concern, it focuses on only a single forecasting date, 1920. In addition, even if the model’s long-run predictions are flawed, it could be a useful predictive tool on intermediate time scales (e.g., 30 y), where unpredictable technological jumps are less likely to make past data unrepresentative of future outcomes. Although 30 y may seem a short forecast horizon for a problem as long-lived as climate change, fully 50% of the value of the SCC is determined by outcomes over this period under some parameterizations of the DICE model (23).

To address these issues, Fig. 2 extends the analysis of Fig. 1, summarizing the model’s predictive performance at each year in the data series, for 30- and 50-y training and confirmation windows. For each year in the period 1900–1980 (1920–1960) the model was trained on the previous 30 (50) y of data, and the estimated model used to forecast GDP for the next 30 (50) y. Although the model performs well in some 30-y periods, in most years the realized growth outcome falls outside of the forecasted interval. Arguably, the model is thus not a successful predictive tool on this shorter time scale, despite less sensitivity to large unpredictable shifts in the technological frontier. For 50-y forecasts, the model performs well post-World War II, but poorly in the prewar period. This shows that the illustration of model performance depicted in Fig. 1 is not exceptional. The 50-y forecasts

---

This finding is dependent on choices of welfare parameters, which in turn affect the social discount rate. All else equal, lower (higher) social discount rates make SCC values more (less) dependent on forecasts of the near future.
further illustrate the sensitivity of growth projections to structural breaks of the kind that followed the war and demonstrate the value of a long historical perspective. Had we only evaluated the model on the most recent 60 y of data we would likely have overestimated its long-run predictive performance.

Our analysis suggests that the version of the neoclassical growth model that DICE relies on could be subject to structural errors on the temporal scales relevant to climate policies. The Ramsey growth model, and more complex models that endogenize the process of technical change, have been profitably applied to a variety of empirical questions in macroeconomics. It is thus important to understand how the use of these models in DICE and other climate applications differs from their more standard empirical applications. Growth models are usually used in empirical applications to explain cross-country differences between the historical growth paths of different countries. In BC-IAMs these models are used to predict the absolute level of global or regional economic output over the coming centuries. When neoclassical growth models are used to explain differences in past outcomes across countries, technical change, in the form of the growth rate of TFP, is formally nothing more than a residual in a linear regression. It is the part of empirical growth data that is not explained by the endogenous factors in the model (i.e., the productivity of capital and labor). If, however, such models are used to make predictions, as in DICE, the future realizations of TFP must also be predicted. This requires us to posit an explicit quantitative model of the evolution of TFP over the coming decades. However, we have no law-like theory of long-run technical change that parallels the predictive successes that have been achieved in the physical sciences (24). This seems unlikely to change in the near future, and more sophisticated models that endogenize the process of technological change also seem unlikely to provide high-powered predictive tools, despite their more nuanced representation of its causes. Just as natural selection explains differences in species' phenotypes without predicting future adaptations, so growth theory has proven to be an insightful tool for explaining the causal determinants of cross-country difference in historical growth outcomes. Prediction, however, is a different matter.

Implications for the Development and Use of BC-IAMs in Quantitative Policy Applications

What can be concluded from this first example of a model confirmation exercise for the economic components of a BC-IAM? Of course, DICE is a single (albeit prominent) example of a BC-IAM, and its implementation of the Ramsey model a single (albeit frequently deployed) representation of the process of long-run economic growth. Our findings are not necessarily representative of how other growth models might fare in similar confirmatory exercises, the structural validity of other BC-IAMs, or the performance of their economic components. Our point, however, is that because most of the economic components of BC-IAMs have not, to our knowledge, been subjected to empirical tests of structural validity using historical data series, we do not know what their empirical status is for quantitative policy applications. Without testing models in hindcasting tasks closely

---

*Empirical tests of endogenous growth models suggest they are often more difficult to reconcile with historical data than simpler neoclassical alternatives (25, 26). Supporting Information, Endogenous Growth Versions of the DICE Model describes our attempt to perform a similar model confirmation exercise on two endogenous growth models recently suggested as alternatives to DICE’s model of TFP growth (27). We find that both models are poorly behaved, being either not specified in a manner amenable to empirical estimation or exhibiting instabilities that cause their predictions to be uncontrolled when their parameters are estimated from historical data.*
related to the uses we wish to put them to today, we cannot
gauge the extent of any possible model errors.

We close with four recommendations for the testing and use of
BC-IAMs in policy applications. First, those components of BC-
IAMS whose structural properties can be meaningfully tested
using historical data should be. Confirmation exercises can build
confidence in model components that perform well historically
and indicate the range of model parameters that needs to be
considered for a given model to have a chance of making sensible
out-of-sample forecasts. If no such parameter range can be
identified from such an exercise, the model’s structural as-
sumptions might need to be revised for the purposes of policy
applications. Calibration (a within-sample exercise) and para-
metric uncertainty quantification techniques are not substitutes
for this procedure, because we need to test the out-of-sample
performance of a model’s structural assumptions. This is not to
say, however, that untested or structurally flawed models cannot
be useful for illustrating qualitative conceptual points about
alternative modeling assumptions. Many BC-IAMs are used
profitably for this purpose in the academic community today, and
none of our points undermines the value of such modeling ex-
ercises if they are interpreted with sufficient carefulness. The
hurdles a model must jump over to qualify for application to
quantitative policy choices should, however, be more demanding.
Second, the economic components of BC-IAMs should be
based on testable structural hypotheses, in so far as this is pos-
able. The confirmation exercise we conducted with the DICE
model was only possible because the model is specified in a self-
contained manner, and with sufficient structural detail to allow it
to be meaningfully compared with data. This is a great virtue of
the DICE model. Although in this case our confirmation exercise
raised the possibility of quantitatively meaningful errors when
applying this model to policy questions, we were at least able to
ask (and partially answer) the question, What risks might we run
by assuming that the world behaves in line with the model? Models
that rely more heavily on external scenarios for key economic
components do not allow for this kind of empirical
testing. If a model component cannot be tested, we cannot
hope to gain confidence in it ex ante, even if it in fact turns out to
perform well ex post. Thus, although a set of exogenous scenarios
could turn out to capture our underlying uncertainty, we can
never estimate what risks we might run by assuming this to be the
case when making decisions today. The practice of making
the assumptions in BC-IAMs testable could help to build confidence
in their outputs and filter out plausible from implausible struc-
tural assumptions.*

Third, policy choices should be based on estimates from many
plausible, structurally distinct, models. As we have noted there are
many important aspects of BC-IAMs that we cannot hope to test
effectively today, because the relevant verification data will only
be realized long after current policies are enacted. There is thus
substantial irreducible uncertainty about some of the core struc-
tural relationships in BC-IAMs. Exploring a wide range of struc-
tural assumptions—not just about overall technological change
and economic growth but also about climate damages and abate-
ment costs—is crucial if we wish the policy prescriptions from
modeling exercises to more accurately reflect the extent of our
uncertainty about the consequences of climate policies. In our
view, and that of others (5, 29), the set of BC-IAMs that are
commonly applied in current policy analysis may underestimate
the risks of inaction on climate change, in part because of a
comparative lack of structural heterogeneity.

Fourth, the decision tools that are used to select policies should
reflect the fact that our models are at best tentative predictive
tools. Modern decision theory has developed a rich suite of tools
for rational decision making under deep uncertainty that allow us
to express our lack of confidence in model output, yet most policy
analysis with BC-IAMs still relies on decision tools that treat the
uncertainty in climate policy as if it were of the same character as
tossing a coin or rolling a die (30). We should instead accept the
limits of our knowledge and use decision tools that fit the pro-
foundly uncertain task at hand.

ACKNOWLEDGMENTS. This work was supported by the Centre for Climate
Change Economics and Policy, funded by the Economic and Social Research
Council, and the Grantham Foundation for the Protection of the En-
vironment (A.M.) and the University College Cork Strategic Research

*Ref. 28 presents a further example of the benefits of hindcasting in a model of US
energy intensity.

1. Oreskes N, Shrader-Frechette K, Belitz K (1994) Verification, validation, and con-
Mitigation of Climate Change, eds Edenhofer O, et al. (Cambridge Univ Press, Cam-
bridge, UK), Chap 6.
change: Grafting gross underestimation of risk onto already narrow science models.
6. Pindyck R (2013) Climate change policy: What do the models tell us? J Econ Lit 51(3):
860–872.
9. Grubb M, Köhler J, Anderson D (2002) Induced technical change in energy and en-
vironmental modeling: Analytical approaches and policy implications. Annu Rev
document: Technical update of the social cost of carbon for regulatory impact
11. Committee on Assessing Approaches to Updating the Social Cost of Carbon; Board
on Environmental Change and Society (2016) Assessment of Approaches to Upgrading
the Social Cost of Carbon: Phase I Report on a Near-Term Update (National Academies
of Sciences, Engineering, and Medicine, Washington, DC), pp 1–74.
from the RICE-2011 model. Working Paper 17540 (National Bureau of Economic Re-
search, Cambridge, MA).
New Haven, CT).
model incorporating the IPCC’s five reasons for concern. Integrated Assess 6(1):19–56.
Princeton).
headwinds. Working Paper 18315 (National Bureau of Economic Research, Cam-
bridge, MA).
574–602.
495–525.
risk: How Nordhaus’ framework supports deep cuts in carbon emissions. Econ J
125(s83):574–620.
Supporting Information

Millner and McDermott 10.1073/pnas.1604121113

An Abstract Definition of Structural Soundness for IAMs

An integrated assessment model is a map \( F \) from initial conditions \( I \), empirical estimates of parameter values \( P \), ethical assumptions \( E \), and policies \( \Pi \) to consequences \( C \). In the training data case, \( F(\Pi; I, P, E) = C \). The subscript \( E \) in \( C_E \) indicates that these are consequences deemed to be relevant according to the ethical assumptions \( E \). The consequence space may be multidimensional and include distributions over, for example, time-dependent streams of consumption. All or some of \( I \), \( P \), and \( C(E) \) may take values in the space of probability distributions over more primitive spaces, or another mathematical structure capable of representing certain knowledge. The policy space may be equally rich; for example, it could include straightforward “open-loop” policies, as well as more sophisticated “feedback” policies that respond to new information when it becomes available. Now let \( RW(\Pi; E) \) be the “real-world” mapping of policy choices to consequences. This map does not depend on initial conditions or parameter values (because these are both fixed in the real world), but it may still take values in the space of probability distributions over primitive consequences, owing to measurement error, missing observations, and so on. In addition, define a “similarity metric” \( d(C, C') \) on the space of consequences. This function should represent the similarity between two consequences. So, for example, if consequences take values in the space of probability distributions over a metric space, \( d(C, C') \) could be the Prohorov metric. We say a model \( F \) is structurally sound on a set of policies \( \Omega \) if

\[
\forall \Pi \in \Omega, d[F(\Pi; I, P, E), \text{RW}(\Pi; E)] < \epsilon,
\]

where \( \epsilon \) captures our tolerance for structural errors on the policy set \( \Omega \). This definition ensures that the structural soundness of a model is always defined relative to its ethical assumptions. Thus, for example, if we adopted an ethical position that assigned zero weight to future generations a model would be considered structurally sound if its short-term predictions were always accurate, even if its long-term predictions were not. Note that in practice we only ever observe the value of \( RW(\Pi; E) \) for one observed policy \( \Pi_0 \). Thus, in its most demanding form (i.e., without auxiliary assumptions that allow plausible counterfactuals to be constructed) the process of model confirmation is restricted to cases where \( \Omega \) is a singleton set. Although this makes it impossible to establish structural soundness on larger sets, the definition may still be fruitfully deployed to show that a given model is not structurally sound on any set \( \Omega \) that contains \( \Pi_0 \). This distinction encapsulates the difference between model confirmation and model validation. A model is confirmed if it has not been shown to be structurally unsound. However, no model can ever be “validated” in an open system (1).

The Model

The economic growth model that DICE employs is a version of the Ramsey neoclassical growth model. At each time \( t \) produce economic output \( Y_t \), using capital \( K_t \) and labor \( L_t \) as factors of production. The aggregate production technology of firms in the economy is a function \( F(K_t, L_t) \), which maps factor inputs to economic output. An amount \( C_t \) of output is consumed, and the remainder is re-invested in firms. Capital is assumed to depreciate at a constant rate \( \delta_K \). The stock of capital \( K_t \) thus evolves according to

\[
K_{t+1} = Y_t(K_t, L_t) + (1 - \delta_K)K_t - C_t.
\]  

[S1]

The DICE model makes several assumptions about functional forms, parameterizations, and data inputs in its implementation of the model. Most important is its assumption about the production technology. \( Y_t(K_t, L_t) \) is assumed to be of Cobb–Douglas form and to include an exogenous time varying TFP multiplier, \( A_t \):

\[
Y_t(K_t, L_t) = A_t K_t^\alpha L_t^{1-\alpha}.
\]  

[S2]

The parameter \( \alpha \) is the capital share of production, which is assumed to be constant and equal to 0.3 in DICE. In our simulations we adopt \( \alpha = 0.25 \), because this is the long-run average capital share in our data (20). Sensitivity experiments on \( \alpha \) suggest that \( \alpha = 0.3 \), provides substantially worse forecasts in our data.

DICE 2013 adopts the following model for the evolution of \( A_t \):

\[
A_t = A_{t-1}(1 + g_t);
\]  

[S3]

\[
g_t = \frac{g_{t-1} + \delta_t}{1 + \delta_t}; g_0 = \text{initial growth rate.}
\]

Training. Once a window of training data is identified, the goal is to estimate the parameters of the growth process for TFP. Throughout the analysis we assume that \( \delta_t = 0 \). This is a conservative assumption, because allowing \( \delta_t \) to be nonzero and estimating both \( g_0 \) and \( \delta_t \) from data yields substantially worse predictions. This is partly due to standard overfitting effects, and partly to the fact that for many training windows the estimated value of \( \delta_t \) is negative, leading to superexponential growth projections (i.e., growth rates that grow exponentially) that far exceed realized values. Thus, our model of the evolution of \( A_t \) is

\[
A_t = (1 + g_0)^t A_0.
\]  

[S4]

The training data consist of a set of time series for \( Y_t, K_t \), and \( L_t \). For each \( t \) in the training window, the inferred value of \( A_t \) is given by

\[
A_t = Y_t / (K_t^\alpha L_t^{1-\alpha}).
\]  

[S5]

The inferred values of \( \{A_t\} \) are then fit with the exponential function (Eq. S4) to obtain 95% confidence intervals on the parameters \( g_0, A_0 \) from the training data. We treat \( A_0 \) as a free parameter because “true” predictive TFP values are measured with error in practice, and to give the model greater flexibility to match the data. This is again a conservative assumption; a model that only estimates \( g_0 \) from the training window and uses the current value of \( A_t \) to generate predictions [i.e., \( A_{t+1} = (1 + g_0)^t A_t \)] fits past data worse and yields narrower confidence bounds and worse predictions out of sample.

Verification. Given initial values of capital and labor in the verification data, and the estimated values of \( g_0, A_0 \) from the training data, the state equation for capital (Eq. S1) is used to forecast a time series for future capital stocks. To do this we require values of \( L_t, C_t \), and \( A_t \). We assume a perfect forecast of labor supply, and thus the actual realized values of \( L_t \) in the verification data are used in the forecasting step. This is a conservative assumption with respect to predictive performance and focuses attention on the model’s representation of technical change.

In the standard formulation of the DICE model consumption \( C_t \) is chosen to optimize an intertemporal utility integral, which represents the welfare of current and future generations. This makes the values of \( C_t \) dependent on the parameters of this utility function, in
particular the elasticity of marginal utility, and the pure rate of time preference. In practice the sequence of savings rates that emerges from this exercise is approximately constant. To avoid having to calibrate the parameters of the welfare function to match past savings data, we assume the savings rate is constant. Thus, consumption $C_t$ is assumed to be a constant fraction $(1 - s)$ of output: $C_t = (1 - s)Y_t$. The savings rate $s$ is set at 22%. This represents a long-run average consistent with historical data and is a commonly used value in the integrated assessment literature (see ref. 17 for justification and discussion of the constant savings rate assumption). Sensitivity experiments suggest that the savings rate needs to be in excess of 85% to cause the verification to fail inside the confidence interval of a model forecast made in 1920, given inferred growth rates of TFP. This is more than three times the average savings rates we observe in the data.

Finally, to specify the values of $A_t$ we use the estimated values of $g_0, A_0$ from the training data and Eq. S4 to generate a forecast of TFP. Once a forecasted time series of capital stocks and TFP is generated, a GDP forecast is obtained by combining the estimated value of $A_t$ with the forecasted value of $K_t$ and the actual value of $L_t$ according to the relationship (Eq. S2).

**The Data**

Our model confirmation exercise is made possible by recent estimates of long-run capital stocks for major world economies by Piketty and Zucman (21). Ideally we would conduct this exercise using global data. However, long-run global estimates of capital stocks do not exist. Instead, we focus our analysis on the United States, given its status since the early 20th century as the global frontier economy (21). In theory, global TFP growth is a function of technical progress in the frontier economy, with other regions “converging” toward the frontier. This is the approach taken by Nordhaus in analysis using the regional version of the DICE IAM (31). He projects technological change for the United States, as the frontier region, and assumes that other countries converge partway to the frontier. As it happens, Piketty and Zucman argue, based on analysis of their historical data, that “as a first approximation, productivity growth is the same everywhere in the rich world” (ref. 21, p.1285). The Piketty-Zucman data are particularly well-suited to our exercise, because they enable us to meaningfully test the long-run predictive performance of the DICE model’s representation of technical change. Whereas official data for the United States only provide capital stock estimates from 1925 (see US Bureau of Economic Analysis Table 1.1: Current-Cost Net Stock of Fixed Assets and Consumer Durable Goods, www.bea.gov/), the Piketty-Zucman data contain annual estimates since 1870.

The specific data series we use for $Y_t, K_t$, and $L_t$ are all taken from the Piketty-Zucman data, available as annual time series from 1870 to 2010 for the United States. The data are available online at gabriel-zucman.eu/files/capitalback/USA.xlsx. For $L_t$, we follow DICE and use total population, without any adjustment for participation rates or other demographic factors. The data series we use for total population of the United States (1870–2010) is from column 17 of Table US1 in ref. 20. For national income, $Y_t$, we use the series in column 3 of Table US1 (expressed in 2010 billions US$). As our measure of $K_t$, we take national wealth estimates from column 9 of Table US6a. The figures in that table are expressed as a percentage of national income, which we convert to a 2010 dollar amount using the $Y_t$ series.

**Discussion**

Over the long run, economic growth is primarily determined by technical progress or the rate of TFP growth (31). Unfortunately, we do not have good predictive theories of how TFP evolves (23, 32). In the economic growth literature, TFP represents the part of growth that remains unexplained after accounting for other inputs and is generally calculated as a residual in growth regressions (see, e.g., ref. 18). The use of a dynamical model of TFP as an essential predictive hypothesis, as in DICE, is therefore slightly nonstandard.

Most climate economy models tend to treat TFP as an exogenous time series (e.g., ref. 14), calibrated to result in a plausible rate of growth over the long run—typically based on extrapolating recent trends. For example, in the original DICE model specification (see ref. 33) the starting value for TFP growth was based on a simple average of estimated TFP growth over a short historical window (1961–1970) before the model start date. This growth rate was then assumed to decline gradually (by half every six decades), roughly in line with projections made in a number of other studies, as surveyed in ref. 34. In more recent incarnations of DICE (see ref. 14) TFP growth was calibrated so as to achieve specified growth rates in consumption per capita.*

A difficulty with extrapolating recent trends in TFP to produce a projection for long-run growth relates to the instability of TFP growth rates over time. In our analysis there is evidence of a structural break in the historical series of TFP growth rates. A Quandt likelihood ratio (QLR) test for the presence of structural breaks rejects the null of stability at the 5% level (the QLR test statistic in 1930 is 4.94; see refs. 35 and 36 for a detailed explanation of the QLR test). These unpredictable structural breaks make long-run forecasting very challenging.

A long-run historical perspective also suggests that using recent growth data as representatives of future trends may be a risky forecasting strategy. The mid to late 20th century seems to have been something of a unique period in the history of economic growth (20, 21). Growth in income per capita was virtually non-existent before 1750. Since then, the frontier economies have experienced a remarkable period of growth, with a pronounced acceleration during the 20th century. For a variety of reasons—notably a slow-down in the pace and real economic benefits of technical innovation—this remarkable growth might not be sustainable (21). More optimistic views hold that we might be on the cusp of the next wave of innovation, driven largely by advances in information technology (37). From a modeling perspective, although it is feasible to represent these possibilities by endogenizing the growth process (38) and explicitly incorporating uncertainty into our representation of economic growth, any given structural model will be a speculative predictive tool, and thus our confidence in quantitative predictions from such models will likely be low.

**Endogenous Growth Versions of the DICE Model**

In this section we report the results of an attempt to perform a similar model confirmation exercise on a modification of the DICE model suggested by Dietz and Stern (27). These authors suggest two possible modifications to the way DICE models technological progress, both of which are designed to endogenize TFP growth. The two models considered are

$$\text{model } 1: A_t = aK_t^b$$

[S6]

$$\text{model } 2: A_t = (1 - \delta_A)A_{t-1} + \gamma_1(I_t)^\gamma_2.$$  

[S7]

Here $A_t$ is the value of TFP at time $t$, $K_t$ is the capital stock, and $I_t$ is the investment flow at time $t$ (given by $sY_t$ in a constant savings rate model). In model 1, $a > 0, b > 0$ are parameters. In model 2, $\delta_A > 0$ is a depreciation rate on $A_t$, and $\gamma_1, \gamma_2 > 0$ are parameters. Both models are designed to capture the idea that knowledge is

*In DICE 2013 $g_0(2015) = 7.9\%$ per 5 y and $\delta_A = 0.6\%$ per 5 y. This specification leads to growth in consumption per capita of 1.89% y from 2010 to 2100 and 1.07% y from 2100 to 2200. See the 2013 DICE user manual (15) for further details.

†The summer 2015 issue of the Journal of Economic Perspectives (Vol. 29, No. 3) contains an illuminating symposium on the possible effects of digital innovation on labor markets and growth.
produced as a “spillover” from capital formation (model 1) or investment (model 2). Model 1 is a modified version of one of the original “first-generation” endogenous growth models (39). A key feature of this model is that if \( a + b > 1 \) (where \( a \approx 0.3 \) is the capital share of production), the model exhibits increasing returns in the capital stock. Model 2 is the authors’ own model, designed to address the fact that \( A_t \) depreciates at the same rather high rate as capital in model 1. In the author’s calibration of model 2, \( \delta_A = 1\% / y \), whereas capital depreciates at a much higher rate of \( 10\% / y \).

We wished to perform a similar model confirmation exercise with these models. Model 2, favored by the authors, is, however, poorly suited to empirical tests. The reason is that because \( \delta_A \) is small, the term \( A_t - (1 - \delta_A)A_{t-1} \approx A_t - A_{t-1} \), which can be negative (this occurs frequently in our data), whereas the term \( \gamma_1(I_t)^{\gamma_2} \) must be positive. The mismatch between the signs of these two terms leads to empirical parameter estimates that are highly unstable, and frequently complex.

Model 1 seems to be a better-behaved model, because both \( A_t \) and \( K_t \) are positive by definition. We have redone our analysis with this model, using training windows to estimate confidence intervals for the two parameters \( a, b \), and then using the coupled state equations for \( A_t \) and \( K_t \) (i.e., Eqs. 51 and 56) to produce forecasts of TFP and GDP. We find that for early training windows the estimation and forecasting steps are well-behaved, although predicted GDP growth is downward-biased (Fig. S1). However, beyond about 1948, the model’s forecasts become highly unstable. The reason is that in later training windows empirical estimates of the upper bound of the 95% CI for parameter \( b \) in model 1 increase substantially (from about 0.3 to about 0.7–0.8). With these large exponents there is a substantial positive feedback effect in the model, with large values of \( A_t \) driving increased productivity, which increases capital formation, which in turn increases the formation of \( A \). When combined with population growth this leads to superexponential growth. For some training windows the model predicts that GDP could multiply by \( 10^{10} \) over 50 y. Thus, we conclude that this model may also not be a well-behaved empirical tool. Indeed, it is well known that models of this kind need not admit stable equilibria (i.e., balanced growth paths) if population is not constant (ref. 18, p. 401). Dietz and Stern (27) calibrate the value of \( b \) to be 0.3; hence, they do not detect the instability we discuss. Their calibration falls outside the 95% confidence interval for \( b \) in our 50-y training windows from 1950 onward.

These exercises reinforce our main conceptual point about the value of “taking models to data” in confirmation exercises. Simply calibrating a model based on “stylized facts” or informed guesswork does not tell us whether the calibrated parameter values are in line with empirical observations, or whether the model’s predictions will be well-behaved when its parameters are estimated from data.

These endogenous growth models are certainly not the only possibilities. More sophisticated “second-generation” models of endogenous growth, including models of directed technical change, Schumpeterian growth, human capital formation, expanding varieties, and so on, could be considered as possible alternatives. These models are interesting and important but would require an entirely different set of data for verification. In particular we would likely require much more detailed long-run data on the specifics of innovation processes. Such data are, to our knowledge, not currently available over comparably long time spans. We hope that our work will stimulate further empirical examination of these models on time scales relevant to climate policy applications.

![Fig. S1. Results of an attempted confirmation exercise with model 1 (Eq. S6) for 50-y training and verification windows. Notice the instability in the model’s predictions that begins around 1948.](image-url)