FORECASTING THE PATH OF U.S. CO_2 EMISSIONS USING STATE-LEVEL INFORMATION

Maximilian Auffhammer and Ralf Steinhauser*

Abstract—We compare the most common reduced-form models used for emissions forecasting, point out shortcomings, and suggest improvements. Using a U.S. state-level panel data set of CO_2 emissions, we test the performance of existing models against a large universe of potential reduced-form models. We find that leading models in the literature, as well as models selected based on an emissions per capita loss measure or different in-sample selection criteria, perform significantly worse compared to the best model chosen based directly on the out-of-sample loss measure defined over aggregate emissions.

I. Introduction

THE possibility of global climate change and its consequences pose a significant environmental threat to humanity over the next century and beyond. The Intergovernmental Panel on Climate Change (IPCC) predicts an increase in globally averaged surface temperatures ranging from 1.1°C to 6.4°C by the end of this century (IPCC, 2007). The main trace gas from anthropogenic sources partially responsible for this warming is CO_2, which accounts for approximately half of the radiative forcing. Its share is predicted to rise to two-thirds over this century (IPCC, 2001a).[1]

Forecasts of future trace gas emissions serve two main purposes. First, emissions of greenhouse gases partially determine their atmospheric concentrations, which are the crucial input to global circulation models used to predict impacts on surface warming and precipitation trends. Second, individual countries use business-as-usual (BAU) forecasts to calculate expected costs of emission reductions in future periods.[2] Since the lion's share of anthropogenic CO_2 emissions in industrialized countries stems from the combustion of fossil fuels, potential reductions in emissions are closely tied to economic activity. A higher level of anticipated emissions in a future period raises the expected costs of emissions reductions relative to a predetermined baseline (for example, the 1990 emissions as in the Kyoto Protocol). Producing optimal forecasts of global and country-level emissions therefore has a direct effect on predicted warming trends as well as the optimal policy decisions made, such as whether to sign and ratify a global climate accord. Further, potentially biased predictions of trends in warming and precipitation due to suboptimal forecasts of emissions will result in biased estimates of the expected benefits from preventing global warming and therefore have an indirect effect on the policy adoption decision.

BAU forecasts of national and global CO_2 emissions are calculated using a variety of modeling approaches, drawing on data disaggregated across sectors or space. The science and engineering literature employs large-scale simulation models (IPCC, 2001b; U.S. Energy Information Administration, 2004). The related branch of the economics literature uses economy-wide input output tables to construct computable general equilibrium models with varying degrees of sectoral detail (Garbaccio, Ho, & Jorgenson, 1999; Böhringer & Welsch, 2004). A less data-intensive approach to forecasting emissions in the economics literature uses reduced-form models based on the environmental Kuznets curve (EKC) hypothesis to construct BAU paths (Schmalensee, Stoker, & Judson, 1998; Holtz-Eakin & Selden, 1995).

Existing applications of these models have three main drawbacks. First, there are *no* published versions of the models that use a given model's ability to predict out of sample in the specification and parameterization process. All existing studies have parameterized the models according to varying measures of in-sample fit, which can be suboptimal when a given model is used for forecasting (Diebold & Mariano, 1995). Second, the chosen models appear to be selected by searching over a narrow space of specifications or parameterizations. This is important, since in-sample selection criteria, such as the Akaike or Schwarz information criteria, are consistent only if the model space contains the true model. Finally, the econometric reduced-form models are usually estimated using per capita emissions, which is equivalent to minimizing a per capita measure of estimation loss. If one is interested in predicting aggregate emissions (Schmalensee et al., 1998; Holtz-Eakin & Selden, 1995), the loss function of interest should be defined over aggregate forecast error.

We make three specific contributions to the literature on carbon emissions forecasting. First, rather than considering only an ad hoc subset, we estimate over 27,000 different reduced-form models arising from possible permutations of a very limited standard set of explanatory variables used in forecasting carbon dioxide emissions at the national level. When we select a "best model" from a forecasting perspective, we use a statistical test of model superiority that accounts for the often neglected effect of data snooping, which may result in selecting a best model by chance. Second, we demonstrate the consequences of selecting forecasting

Received for publication September 21, 2008. Revision accepted for publication June 15, 2010.

* Auffhammer: University of California, Berkeley and NBER; Steinhauser: Australian National University.

We thank the editor, Mark Watson, two anonymous referees, Peter Berck, Richard Carson, Jeffrey LaFrance, Andrew Patton, and seminar participants at Cornell University and the 2008 EAERE conference for helpful suggestions. All remaining errors are our own.

[1] The main greenhouse gases in addition to CO_2 are methane (CH_4), nitrogen oxides (NO_x), nitrous oxide (N_2O), ozone (O_3), and halocarbons such as $CFCL_3$ and CF_2CL_2.

[2] Business-as-usual forecasts, which are sometimes referred to as baseline forecasts, are predictions of future realizations for sequences that are potentially affected by policies, in the absence of additional policies. They serve as a point of comparison for predictions assuming policy intervention. One example relevant to this paper is the business-as-usual forecasts of U.S. CO_2, emissions in the absence of federal climate regulation, which can then be compared to forecasts under more or less aggressive federal climate policy scenarios.

The Review of Economics and Statistics, February 2012, 94(1): 172–185

models based on different objective performance measures varying in the validation horizon (in sample versus out of sample) and aggregation over which the forecast loss function is defined (per capita versus aggregate emissions). Finally, we compare our forecasts to those based on classic benchmark specifications found in the literature.

In our analysis, we make use of a state-level panel data set for U.S. CO_2 emissions covering the years 1960 to 2001. The United States has been the leading emitter of CO_2 since 1870 (Marland, Boden, & Andres, 2004). It has recently been overtaken by the People's Republic of China as the leading emitter (Auffhammer & Carson, 2008). Constructing BAU forecasts for U.S. CO_2 emissions is of crucial importance, since the United States has neither ratified the Kyoto Protocol nor independently implemented a federal program to reduce emissions. Previous administrations have repeatedly cited the high anticipated costs of carbon reductions as a reason for not joining a global accord. If suboptimal forecast models falsely predict high emissions, this would result in artificially high expectations of abatement costs at a future date. Our paper outlines a rather simple strategy to construct optimal U.S. CO_2 emissions forecasts from a forecaster's perspective. A growing literature shows that forecasting the aggregate by exploring variation at a smaller geographic scale may result in superior forecasts (Marcellino, Stock, & Watson, 2003; Giacomini & Granger, 2004). We follow this approach by forecasting emissions at the state level and aggregating them up to the national level to exploit these forecasting efficiency gains.

To preview the results, we find that benchmark models from the literature are outperformed by 25% of the models in our universe. We show that the choice of performance measure is consequential and argue the need for a shift in the emission forecasting literature toward criteria based on out-of-sample ability to predict total emissions versus the standard approach of selecting per capita models based on in-sample fit. The best model selected according to loss defined over per capita emissions predicts 2011 emissions to be 4% lower compared to the best model based on loss defined over aggregate emissions. The difference is equivalent to 5% of 1990 U.S. emission levels. To put this number in perspective, under the Kyoto Protocol, the United States made a nonbinding promise of emission reductions of 7% relative to 1990.[3] The carbon emission forecast of our best model predicts emissions 100 million tons of carbon lower than the average of studies from the IPCC's *Special Report on Emissions Scenarios* for the year 2010, suggesting that these scenarios may on average overstate the 2010 level of emissions.

The next section provides a brief review of the literature. Section III discusses the data, model universe, and model selection criteria. Section IV provides estimation results and the data snooping tests. Section V concludes.

II. Background and Literature Review

There are two distinctly different approaches to modeling emissions of CO_2. The first is a structural general or partial equilibrium modeling approach in which an often sizable set of free parameters is fixed by judgment and calibration. The second approach employs reduced-form econometric models where a small number of parameters is calibrated to historical data solely based on the goodness of model fit.

Structural modeling is the predominant approach in the natural science and engineering literature. The organizing framework of these models is based on the $I = P \cdot A \cdot T$ identity (Ehrlich & Holdren, 1971). This identity decomposes **I**mpact (emissions) into **P**opulation, **A**ffluence (per capita GDP), and a **T**echnology index. IPAT models imply that emissions increase monotonically in population and affluence and decrease with beneficial technological progress. The more recent engineering literature has focused on modeling the technological change component of the IPAT model. This has involved fine-tuning structural parameters to accurately emulate real data. The most important examples of this class of modeling underlie the IPCC's (2001b) *Special Report on Emission Scenarios* (SRES). The six official simulation models used to produce emission scenarios are the AIM, ASF, IMAGE, MESSAGE, MARIA, and MiniCAM models. These complex simulation models link socioeconomic scenarios to energy-economic and land equilibrium models to arrive at regional emission scenarios, which are then aggregated to a global emissions trajectory for each scenario. Another example of this type of model at the national level is the Energy Information Administration's NEMS model.[4] The IPCC, for political reasons, neither publishes country-level forecasts nor evaluates its own aggregate scenarios' forecast performance.

Structural models are also found in the economics literature. Computable general equilibrium (CGE) models decompose variation in emissions at the sector level by making use of nationally aggregated input-output matrices. A large literature addresses using CGE models to predict carbon emissions for developed and developing countries (Böhringer, Conrad, & Löschel, 2003). This approach to modeling emissions is popular in policy circles, since one can easily simulate the impacts of different policy instruments and shocks to the economy on resulting changes in emissions. These models, while often used to draw out of sample predictions, are not forecasting models since they are not calibrated according to their ability to predict out of sample. Further, these models require a large amount of data, which in many countries are provided at very infrequent intervals (China's input-output tables are provided every five years, for example).

[3] The United States never ratified the Kyoto Protocol and is therefore not bound to engage in these emissions reductions.

[4] The EIA uses the NEMS model to forecast the national energy system and CO_2 emissions, which are published in the "Annual Energy Outlook." Since these forecasts are provided by an agency of the U.S. government, they are considered the official forecasts. The EIA conducts an annual out-of-sample evaluation of its forecasts, but it is not used to reparameterize the NEMS model. For a discussion of the NEMS model forecast evaluation, see O'Neill and Desai (2005) or Auffhammer (2007).

The second branch of the economics literature has focused on reduced-form models. The econometric literature on forecasting emissions is largely based on early work by Grossman and Krueger (1993) and Selden and Song (1994), who look at the in-sample relationship between air pollutants and income. The focus of this literature is the empirical finding of an inverse-U relationship between emissions and ambient concentrations of pollutants and per capita income, which is known as the environmental Kuznets curve. The existence of such a relationship has been at the center of the debate surrounding trends in emissions of local and global pollutants from developing and developed countries alike.

The main criticisms of this model, as Copeland and Taylor (2004) and Arrow et al. (1995) point out, is that this reduced-form specification does not separate the income effect from other factors driving emissions. Recent work by Millimet, List, and Stengos (2003) and Harbaugh, Levinson, and Wilson (2002) casts doubt on the robustness of the EKC specification for local air pollutants. The empirical evidence is mixed on whether an in-sample turning point exists for the odorless and invisible gas CO_2 (Lieb, 2004). Aldy (2005) correctly points out that the existence of a negative marginal propensity to emit (MPE) carbon at the highest levels of income has large consequences for out-of-sample carbon forecasts.

Schmalensee et al. (1998) use a flexible version of the environmental Kuznets curve specification to forecast emissions of CO_2 out of sample. Holtz-Eakin & Selden (1995) before them used a simple quadratic income term to implement such an inverse-U relationship for CO_2 emissions. Both papers use in-sample fit to select their forecasting model, which may lead to suboptimal out-of-sample performance (McCracken & West, 2004). Both use the same source of data, although the latter observe emissions over a shorter time period. Holtz-Eakin and Selden (1995) found a diminishing MPE only at the highest levels of income, which is a finding consistent with the most recent studies by Vollebergh, Melenberg, and Dijkgraaf (2009) and Azomahou, Laisney, and Van (2006). Schmalensee et al. (1998), however, find evidence of a negative MPE.

The major advantage of the reduced-form models from a practical perspective is that they have lower data requirements, which allows the use of longer time series and facilitates the analysis for countries where the structural approach is infeasible. Most important, they avoid the need of their structural counterparts for a large number of parametric assumptions to be made. The difference between the structural and reduced-form approach to forecasting emissions is similar to the debate on macroeconomic forecasting from the 1970s and 1980s. The outcome from that debate was an abandonment of large-scale structural models in favor of much simpler econometric models due to their ability to better predict series of interest out of sample (Wallis, 1989).

The literature on econometric forecasting model selection can be divided into three approaches. The first and most commonly practiced approach is a sequential model selection approach, by which one starts with a general unrestricted model (GUM) based on the largest set of potential regressors and then selects the "best" forecasting model by sequential testing of zero parameter restrictions. In order to circumvent the issue of path dependence, encompassing tests and model selection criteria are applied to select from competing models in order to obtain a parsimonious representation. Specification of the GUM is a crucial step in this approach. If relevant predictors are excluded from the GUM, the chosen forecasting model may result in suboptimal forecasts.

An alternate approach to constructing forecast models is the diffusion index approach of Stock and Watson (2002). If one has a large number of relevant predictors relative to the length of the time series, one constructs principal components from the space of covariates in a first stage, which are then included with lags of the dependent variable in a second stage. Model selection happens by the use of an information criterion. Finally, one could apply Bayesian shrinkage estimation to a most general model and shrink the coefficients toward 0.

Regardless of the adopted approach to model selection, the fact that one observes only a single realization of any time series means a danger that the observed predictive power of the chosen model may be due to chance rather than true forecasting ability of the model. An additional problem is that in practice, most specification searches in practice are not systematic and cannot hope to be comprehensive. This issue, commonly referred to as data snooping, describes any situation in which data are used repeatedly for inference or model selection, but the reuse of the data is not accounted for in inference tests.[5] The reason for oversight of this issue in applied studies was the lack of an easily implementable and broadly applicable way of accounting for the impact of specification searches on inference tests. We make use of a recent, generally applicable method from the financial econometrics literature for testing the null hypothesis that the best model encountered during a specification search has no predictive superiority over a benchmark model (White, 2000; Hansen, 2005).

III. Model Selection and Data

A. *The Model Universe*

The potential set of variables that drive the emissions of CO_2 is very large. The literature on modeling emissions using reduced-form models has focused on a small subset of these variables (Schmalensee et al., 1998; Holtz-Eakin & Selden, 1995; Yang & Schneider, 1998). We define our model universe over this modest set of standard variables used in the literature. Specifically, we have collected state-level data for income, population density, and several categorical variables. The general model considered in this paper takes two forms,

[5] The problem of data snooping has been long understood and was pointed out by Cowles (1933) and Leamer (1978). It has been brought to wide attention by Lo and MacKinley (1990).

which vary by the information set considered. The first set of forecasts, which we call indirect forecasts for the remainder of the paper, are based on the following equation:

$$\begin{aligned} c_{i,t} = {} & \rho_i c_{i,t-1} + f(\text{incomepc}_{i,t}) + g(\text{pdens}_{i,t}) \\ & + s_i + \gamma_t + \varepsilon_{i,t}, \end{aligned} \tag{1}$$

where $c_{i,t}$ are per capita carbon emissions for state i in year t, $\text{incomepc}_{i,t}$ is per capita real personal income, $\text{pdens}_{i,t}$ is population density, s_i is a state fixed effect, and γ_t is a year fixed effect. $f(\cdot)$ and $g(\cdot)$ are flexible functional forms, implemented as either higher-order polynomials or splines. $\varepsilon_{i,t}$ is assumed to be a stationary ergodic error term. The models based on equation (1) allow a lag of the dependent variable, yet the right-hand-side variables enter contemporaneously, which will require forecasts of the right-hand-side variables for the future period of interest $(t + \tau)$.[6]

The second set of forecasts, which we call direct forecasts for the remainder of the paper, are based on the following equation:

$$\begin{aligned} c_{i,t} = {} & \rho_i c_{i,t-\tau} + f(\text{incomepc}_{i,t-\tau}) + g(\text{pdens}_{i,t-\tau}) \\ & + s_i + \gamma_t + \varepsilon_{i,t}. \end{aligned} \tag{2}$$

Models based on this general specification differ from specification 1 in that all information on the additional covariates used to predict emissions in period $t + \tau$ is based on information available in period t. This approach therefore does not require us to make forecasts of any right-hand-side variables.

Starting with these admittedly basic two general models, we sequentially impose restrictions to arrive at a large but finite number of specifications. Table 1 gives an overview of all the variations of equations (1) and (2) that we consider. The unique possible permutations produce a model universe of 27,216 different specifications for this limited set of explanatory variables.[7] The functional form for income $f(\cdot)$ varies from a linear income term up to a fifth-order polynomial, as well as a spline on income with the number of segments varying from three to ten. The functional form of population density $g(\cdot)$ varies from a linear to a quadratic polynomial. In order to model shocks common to all states for a given year, which include technological change, we include year fixed effects. A more parsimonious approach to proxy for technical change is to include a time trend instead of year fixed effects.[8] We allow logarithmic as well as linear trends, both of which are found in the literature. The set of specifications without time fixed effects has the advantage of needing to estimate many fewer parameters. However, these specifications do not control for year-specific shocks common to all states. We therefore allow the inclusion of high-energy-price regimes called crisis dummies in these specifications. Further, in specifications without state fixed effects, we can allow the inclusion of dummies proxying for relevant differences in unobservables across states, such as whether a state is oil or gas producing, coal producing, or located at the coast.

Table 1.—Summary of Variations That Generate the Model Universe for Indirect and Direct Forecasts

Variations of main variables
Income per capita ($incomepc_{i,t}/incomepc_{i,t-\tau}$) up to the fifth-order polynomial
Population density ($pdens_{i,t}/pdens_{i,t-\tau}$) up to the second-order polynomial
Income splines with 3 to 10 segments
Variations addressing temporal and spatial heterogeneity
State fixed effects or state dummies for coastal, oil or gas producing, and coal producing
Year fixed effects
Linear or logarithmic time trend
Energy crisis dummies (1973–1975, 1979–1981, 1990–1991)
Further variations
Levels or logs
Including regressor lags ($pdens_{i,t-1}/pdens_{i,t-\tau-1}$ and $incomepc_{i,t-1}/incomepc_{i,t-\tau-1}$)
Including lagged dependent variables ($c_{i,t-1}/c_{i,t-\tau}$)

Models (1) and (2) are dynamic models, which allow lagged emissions to affect current emissions. Emissions of CO_2 originate from a durable capital stock, which turns over slowly. A static model imposes the implicit restriction that current income and population are the only factors driving emissions. Houthakker and Taylor (1970) outline a simple model in energy demand, where the partial adjustment of a durable capital stock results in lagged emissions affecting current emissions.[9] We include state-specific lagged per capita emissions, which proxy for this more flexible capital adjustment process. We also allow a distributed lag process

[6] Details on how the indirect forecasts are computed are given in section IIIE.

[7] There are two reasons that we did not include further explanatory variables. First, for the indirect forecasts, each additional regressor requires projections to conduct a true out-of-sample forecasting exercise. We obtained quasi-official projections for population, but as we show is the case for income, it would be difficult to obtain such projections for other variables like oil prices or heating degree days that do have an influence on carbon emissions. Forecasting regressors out of sample adds prediction error in models where they are included, thus making it potentially less likely that those models are selected. Second, our model universe based on the sparse set of variables is still manageable computationally, but large enough to demonstrate the core findings of this paper. It is also in the spirit of a reduced-form model approach with its lower data requirements to use a limited set of covariates.

[8] In the forecasting context, year fixed effects are problematic, since they need to be forecast out of sample and therefore add another layer of estimation uncertainty. Models in our universe with year fixed effects vary by the year (1971–1973) in which we allow a structural break to occur when we estimate a time trend around the fixed-effect coefficients to get out-of-sample forecasts. See section IIIE.

[9] As Nickell (1981) points out, the inclusion of a pooled lagged dependent variable in fixed-effects models is problematic and leads to biased coefficient estimates. Judson and Owen (1999) use Monte Carlo techniques to show that a least-squares dummy variables (LSDV) in this setting outperforms the alternative GMM estimator. We estimate all models, which include a lagged dependent variable, using the simple LSDV estimator. The previous literature has not explored heterogeneity in the lag coefficient, which allows state-specific differences in the speed of adjustment. We compared otherwise identical models with pooled and heterogeneous lag coefficients. The heterogeneous models for our sample on average show a 10% improvement in aggregate MSFE. This finding is, of course, sample specific. Since the coefficients in these forecasting models have no causal interpretation, concerns about bias are mute.

TABLE 2.—IN-SAMPLE AND OUT-OF-SAMPLE CRITERIA USED FOR MODEL SELECTION

In-sample	
Akaike information criterion (AIC)	$\ln(\sum_{i=1}^{50} \sum_{t=1960}^{2001} \frac{e_{it}^2}{n}) + \frac{2k}{n}$
Schwarz information criterion (SIC)	$\ln(\sum_{i=1}^{50} \sum_{t=1960}^{2001} \frac{e_{it}^2}{n}) + \frac{k}{n} \ln(n)$
R^2	$1 - \frac{\sum_{i=1}^{50} \sum_{t=1960}^{2001} e_{it}^2}{\sum_{i=1}^{50} \sum_{t=1960}^{2001} (y_i - \bar{y})^2}$
$\bar{R}^2$	$1 - \frac{n-1}{n-k}(1 - R^2)$
Out-of-sample	
Per capita mean square forecast error	$\frac{1}{n} \sum_{t=2001-\tau}^{2001-(n-1)-\tau} \sum_{i=1}^{50} (c_{i,t+\tau} - \hat{c}_{i,t+\tau})^2$
Aggregate mean square forecast error	$\frac{1}{n} \sum_{t=2001-\tau}^{2001-(n-1)-\tau} \sum_{i=1}^{50} (pop_{i,t+\tau} \cdot c_{i,t+\tau} - pop_{i,t+\tau} \cdot \hat{c}_{i,t+\tau})^2$

of income and population density, since these factors may take several periods to affect emissions. All of these possible combinations are included in levels and logarithmic form to allow a multiplicative data-generating process of the regression equation.

B. *Model Selection*

The model selection criterion is the crucial factor in the search for an optimal forecasting model. Existing studies in this literature select models using the R^2, Akaike, or Schwarz information criteria or by performing batteries of joint significance tests on the vector of estimated parameters to arrive at a "best" specification. The main drawback of this strategy is that the forecasting model is selected based on in-sample fit, when the stated goal is to predict emissions out of sample. If the goal is to choose a model with superior out-of-sample predictive ability, one should conduct an out-of-sample prediction experiment to see which model minimizes the cost from out-of-sample forecast error. An explicit statement of the different selection criteria helps to clarify this notion. Anticipating our empirical application, in-sample fitting criteria are based on the sample equivalent of the population disturbance (ε_{it}) for state i at time t from the general unrestricted model. We denote its sample counterpart e_{it} and call it the residual. Table 2 shows the formulas for calculating different standard measures of in-sample fit.

Rather than selecting a model specification that fits best in sample, we conduct the following out-of-sample prediction experiment. In order to compare models based on predictive ability τ years out of sample, we use all information up to period t and predict carbon emissions in period $t + \tau$. In the case of indirect forecasts, we need projections of all right-hand-side variables used in the different model specifications.[10] These are then used together with the coefficient estimates from a regression using the sample up to year t to calculate the indirect models' prediction for the year $t + \tau$.[11] For direct forecasts we do not need projections of the regressors and can calculate the prediction for the year $t + \tau$ using lagged variables. Now we can compare the model's prediction with the actual emissions for state i in year $t + \tau$ and can calculate the forecast error $(c_{i,t+\tau} - \hat{c}_{i,t+\tau})$. We then square the forecast errors and sum over the fifty U.S. states to get the model's forecast error for $t + \tau$. We repeat this procedure for n years starting with $t = 2001 - \tau$ moving backward through time. We now calculate the average over the n periods to get the model's mean square forecast error (MSFE) as defined at the bottom of table 2. This prediction experiment gets us an out-of-sample predictive ability measure for each model, which we can use as a selection criterion for how well models will predict τ periods into the future. In our experiment, we chose both n and τ equal to 10. The choice for $\tau = 10$ was made for two reasons. First, this time span is equivalent to the number of years between the signing of the Kyoto Protocol and the year prior to the beginning of the first commitment period. The second reason is the trade-off between the length of the forecast horizon τ and the number of repeated out-of-sample forecasts n that we can do in the prediction experiment given the limited number of years in the data set.[12] It would be desirable to have a larger number of repeated forecasts n for each model in order to increase the external validity of the prediction experiment and to increase the forecast horizon if one is interested in doing forecasts more than ten years into the future. We acknowledge the fact that one might potentially get a different "best" model if the time period used to evaluate the predictions changed significantly. We cannot meaningfully test for this, given the short time series.

The bottom of table 2 shows that the loss function can be defined over the forecast error for per capita emissions or aggregate emissions. In the latter case, state-level per capita values are multiplied by the state's population to get aggregate emissions before the prediction error is calculated. This

[10] In section IIIE, we discuss in detail how we get the projections for the different explanatory variables.

[11] This calculation is simple for all regressions in the universe estimated in levels. When we estimate the state-level emissions in logs, we need to transform the predicted log per capita emissions back to levels using the exponential function before we compare them to the actual emission using the formulas of table 2. For the logarithmic specification, we use the Goldberger (1968) correction to get our point forecasts in levels.

[12] Even with the given choice of n and τ of 10, we are stretching the data, leaving only 11 years of in-sample data for the earliest model forecasts in the prediction experiments for direct models, which use lagged variables.

has the effect of more heavily weighting errors in the populous states such as California instead of giving a large weight to errors in sparsely populated high per capita emitters like Wyoming or North Dakota. Model selection based on a loss function defined over aggregate U.S. emissions will result in a model that most accurately predicts total U.S. carbon emissions, which is the stated goal.

C. The Benchmark Models

We chose a set of benchmark models that fulfill two important purposes. First, they are used to contrast the point forecasts of the best-performing model from our search with standard specifications from the literature. Second, we use the benchmarks in order to formally test the predictive superiority of benchmarks against the best model from our model universe using the Hansen (2005) reality check bootstrap test (RC) data snooper method. This method compares the best-performing forecasting model from the given model universe to a benchmark (White, 2000; Hansen, 2005). In the original application, Sullivan, Timmermann, and White (1999) compare a large number of technical trading rules to the benchmark of holding cash. The benchmark works as an orientation that makes the performance of the best model quantifiable. We choose two highly cited specifications of the reduced-form literature (Schmalensee et al., 1998; Holtz-Eakin & Selden, 1995) and one basic decomposition model from the structural literature (Yang & Schneider, 1998). Schmalensee et al. (1998) propose the following specification:

$$\ln(c_{it}) = s_i + \gamma_t + F(\ln(\text{incomepc}_{it})) + \epsilon_{it},$$

where the variables are the same as before and $F(\cdot)$ represents a piecewise linear function with ten segments.[13] The model specification that Holtz-Eakin & Selden (1995) used is

$$\begin{aligned}\ln(c_{it}) = s_i + \gamma_t &+ \alpha_1 \ln(\text{incomepc}_{it}) \\ &+ \alpha_2(\ln(\text{incomepc}_{it}))^2 + \epsilon_{it}.\end{aligned}$$

This model corresponds to the most traditional EKC specification.[14] As a last benchmark, we use the nonstochastic structural identity of Yang & Schneider (1998):[15]

$$\begin{aligned}carbon_{it} &= population_{it} \\ &\times \frac{income_{it}}{capita_{it}} \times \frac{energy_{it}}{income_{it}} \times \frac{carbon_{it}}{energy_{it}}.\end{aligned}$$

D. Data

Blasing, Broniak, and Marland (2004) provide a data set of CO$_2$ emissions for the fifty states and Washington, D.C., for the years 1960 to 2001, which results in a balanced panel of 2,100 observations.[16] This is the longest and most complete CO$_2$ emissions data set for a single country at a subnational level of aggregation. They used consumption data for coal, petroleum, and natural gas from the EIA State Energy Data Report to calculate carbon emissions.[17] The data do not account for carbon oxidized during gas flaring or from the calcining of limestone during manufacture of cement or for carbon from bunker fuels.[18] Emissions are reported in million metric tons of carbon. We test the panel of per capita emissions for a unit root and reject the null of a unit root using a Levin-Lin-Chu test and the Im-Pesaran-Shin test at the 1% level. We therefore do not include any specifications in differences.

Income data are taken from the U.S. Bureau of Economic Analysis. We deflate personal income by state into 2001 U.S. dollars. Population data are taken from Blasing et al. (2004). Data on land area of each state were obtained from the U.S. Department of Commerce. We collected qualitative variables for whether a state is located at the coast or is oil/gas or coal producing. Finally, we construct three dummies for the energy crises (1973–1975, 1979–1981, and 1990 to 1991) to reflect their temporary shocks on the system. We err on the side of including an additional year after the recovery of oil prices, as the immediate effects of the shock are still echoing. As Marcellino et al. (2003) point out, if there is a sufficient degree of heterogeneity in state-level series, constructing forecasts of the aggregate by summing up state-level forecasts may result in improved forecasts over forecasting the aggregate directly.

Figure 1 displays per capita emissions for 1960 and 2000 and their growth rates for all fifty states and Washington, D.C. The series display tremendous cross-sectional as well as time series variability. The state-level income series display a similar degree of variability along both dimensions. Average per capita income over the sample period and states is \$21,289, with a standard deviation of \$5,668. In 2001, the poorest state, Mississippi, had an average income of \$21,595, which is only 50% of that of the richest state, Connecticut, with \$42,657. By using data for a single country that are collected using consistent definitions and procedures, we avoid

[13] Schmalensee et al. (1998) use four model specifications. We adopt the specification they refer to as ten-segment income spline and linear trend model as our benchmark. To get out-of-sample forecasts of the year fixed effects, we follow Schmalensee et al. (1998) and regress the year fixed effects on a two-piece spline function that allows a structural break in the time trend at 1970: $\gamma_t = \alpha_0 + \alpha_1 t + \alpha_2(t - 1970) \cdot \mathbf{1}_{\{t \geq 1970\}}$.

[14] Holtz-Eakin & Selden (1995) use the last in-sample period time fixed effect coefficient as the time dummy in their projections.

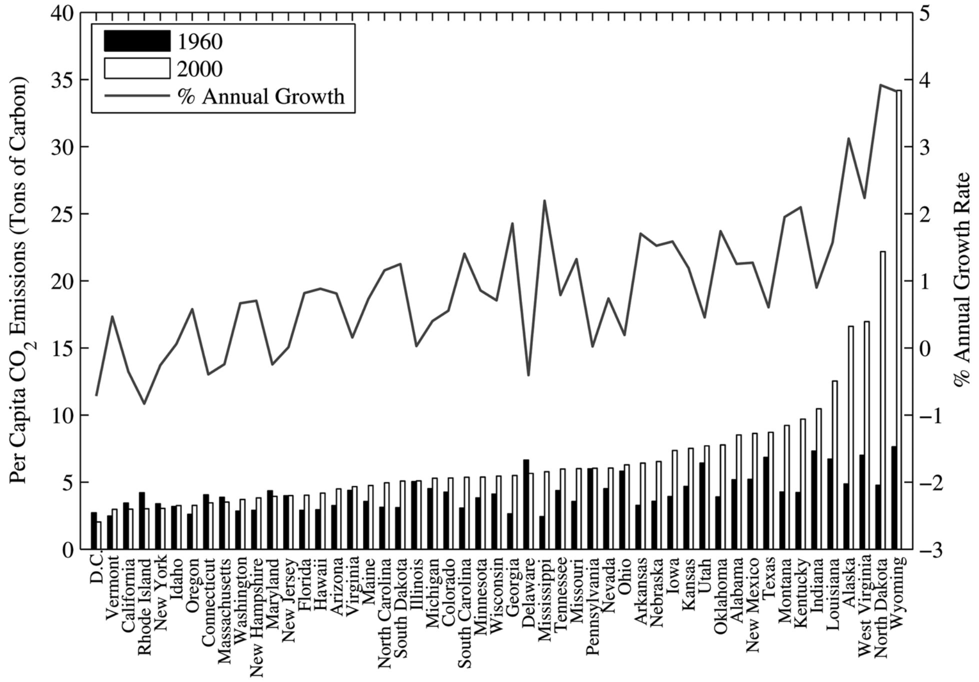
[15] This identity can be thought of as $carbon_{it} = population_{it} \times GDPpercapita \times EnergyIntensity \times CarbonIntensity$. To obtain forecasts for CO$_2$ emissions using this decomposition, one multiplies the base-year emissions with the estimated growth rates of the four factors of the identity. We therefore needed forecasts of energy and carbon intensity in addition to population and income predictions. For energy and carbon intensity, we followed Yang and Schneider (1998) and used predictions for developed countries based on the IPCC (1992) energy production data.

[16] In our empirical analysis, we will concentrate on the fifty U.S. states and omit Washington, D.C.

[17] For petroleum, the data include energy production and transportation in each state, as well as oxidized carbon emissions from other end uses such as the production of plastics, fabrics, or lubricants.

[18] Those neglected sources together account for approximately 4% of the total carbon emissions. See Holtz-Eakin & Selden (1995).

FIGURE 1.—PER CAPITA CO_2 EMISSIONS AND ANNUAL GROWTH RATE, BY STATE



The bars show state per capita emissions for 1960 and 2000. The line shows the average annual growth rate in per capita emissions for all fifty states and Washington, D.C., between 1960 and 2000. States are ordered by their 2000 per capita emissions.

concerns that potential nonlinearities of emissions in income are due to lower measurement error in wealthier countries, which may result in an "artificial" EKC.

E. *Explanatory Variable Projections*

Projections of all right-hand-side variables are required for the indirect forecasts based on equation (1). The required projections of the covariates in our experiment use all available information up to period t to make a forecast of the covariate for period $t + \tau$. Here we require only forecasts of state-level per capita income and population density, yet indirect models using a more comprehensive set of covariates would require forecasts of these additional variables.

There exist several state-level population projections made by the U.S. Census Bureau over the course of the past 25 years that we can use to calculate the projected population density variable. The state population projections we use for forecasts into the future beyond 2001 are based on the Census Bureau, Population Division projections.[19] The forecasts use information up to 2000 Census to make population projections up to 2030. The earlier Census Bureau population projections by Wetrogan (1983) are used for our prediction experiments. These projections use all available information up to 1980 to make population projections for the following twenty years.[20]

Unlike the data for population, we were unable to find suitable historic state-level income projections for the United States ten years out of sample. Thus, for each state, we generate projections using a method similar to that of Auffhammer & Carson (2008), which implicitly controls for the correlation between population and income. We assume that the income growth rate ξ_t and population growth rate ϕ_t for a given state are jointly distributed as $f(\xi_t, \phi_t) \sim N_2(\mu_\xi, \mu_\phi, \sigma_\xi^2, \sigma_\phi^2, \rho)$ and can be characterized in and out of sample by this bivariate normal distribution. The distribution is parameterized by using the in-sample estimated mean and standard deviation of the population growth rate as well as its correlation coefficient with income growth—$\hat{\mu}_\phi$, $\hat{\sigma}_\phi$, and $\hat{\rho}$, respectively. $\hat{\mu}_\xi$ and $\hat{\sigma}_\xi^2$ are the in-sample mean income growth rate and its variance at the state level.

In order to obtain a value for $\xi_{t+\tau}$ given a value of $\phi_{t+\tau}$ from the population projections, we use the expected value of the conditional marginal distribution $f_\phi(\xi_{t+\tau})$. Using the methodology above, we construct an income series that covaries with population growth but also has a random component based on the variance of the historical income growth rate.

Projections of the income and population series are sufficient to conduct out-of-sample forecasts for all indirect models in our universe. For indirect models without lagged

[19] The State Interim Population Projections are published on http://www.census.gov/population/www/projections/projectionsagesex.html. We use a piecewise cubic interpolation to get the missing years 2002–2004.

[20] We use a piecewise cubic interpolation to get the missing years between 1990 and 2000.

dependent variables, calculating the forecasts is straightforward. For the indirect models, which contain a lagged dependent variable on the right hand side, we calculate carbon emissions in period $t + \tau$ using the expression $\hat{c}_{i,t+\tau} = \sum_{k=0}^{\tau-1} x_{i,t+\tau-k}\hat{\beta}\hat{\rho}_i^k + c_{i,t}\hat{\rho}_i^\tau$.

Finally different time and space covariates are included in most models. The inclusion of the time-invariant state fixed effects and state dummies for coastal, oil/gas and coal in the out-of-sample forecast is straightforward. For models that include a linear or logarithmic time trend, we continue these trends into the out-of-sample periods. Models that include year fixed effects create more of a challenge, since one does not know the value of the year fixed effect for future periods. Holtz-Eakin & Selden (1995) set the time fixed effect equal to the last in the sample year, while Schmalensee et al. (1998) attempt to forecast them out of sample, as discussed in note 12. Following the latter approach, we examined a variety of specifications forecasting the time fixed effects out of sample by allowing the breaking point to vary:

$$\hat{\gamma}_t = \alpha_0 + \alpha_1 t^{<B} + \alpha_3 t^{\geq B}. \tag{3}$$

This is a regression of the estimated in-sample time fixed effects on an intercept and a linear time trend, which is allowed to break at year B. The year for the structural break B varies in our model universe from 1971 to 1973 for different model specifications.

F. The Reality Check Bootstrap Test

Once we have chosen the best-performing model based on any of the model selection strategies outlined in section IIIB, we would like to know with some confidence that the model encountered in our specification search has predictive superiority over, for example, our benchmark models. Traditional out-of-sample prediction tests such as Diebold and Mariano (1995) and West (1996) allow for the comparison of two competing models under different sets of assumptions. However, these and other traditional predictive ability tests ignore dependence between the results for different forecasting models. Hansen (2005) provides a test that incorporates the dependence of results across forecasting models into a comprehensive bootstrap-based test. We improves on the earlier technique provided by White (2000), which assigns too much weight to poorly performing models with high variances. This is problematic in applied studies such as ours, where the range of models spans many such poorly performing models.

The so-called reality check bootstrap Test (RC) tests the null hypothesis that the best model encountered in the specification search has no predictive superiority over a given benchmark model, taking the initial search over models into account.[21] As such, it accounts for the possibility that the best-performing model is selected by chance. To obtain the distribution of the test statistic, which is based on the relative performance of the benchmark to the best model, we use a bootstrap approach incorporating all models. The resulting *p*-value provides an objective measure of the extent to which the apparently good results of the best model accord with the sampling variation of the searched universe (White, 2000).

IV. Results

A. Out-of-Sample Prediction Experiment

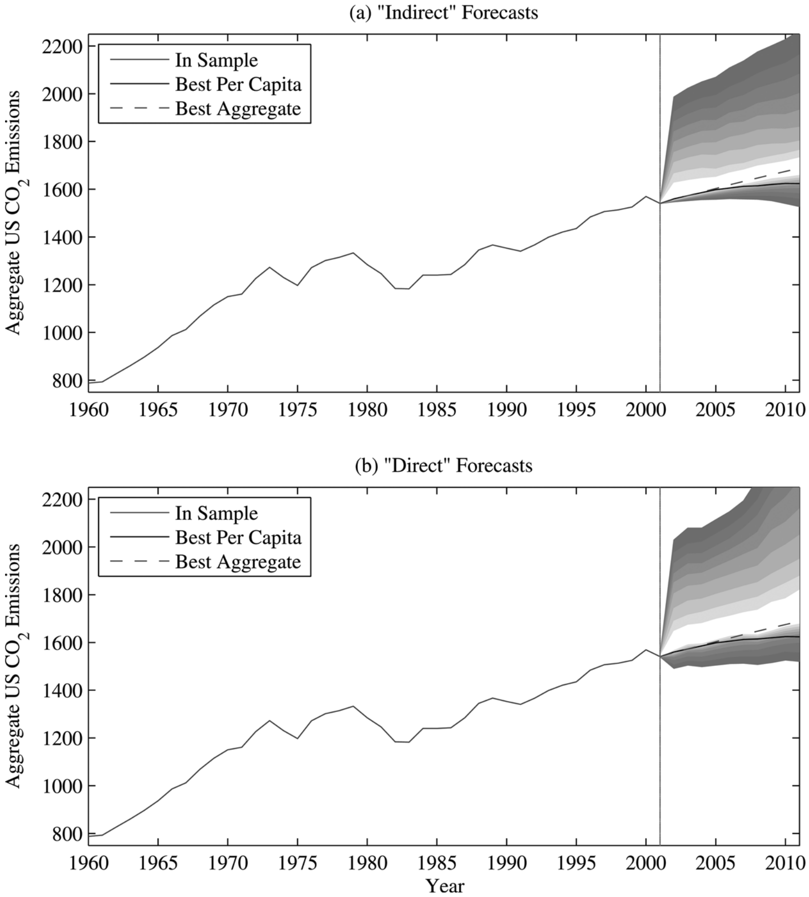
We calculate the out-of-sample MSFE for all 27,216 models in our universe according to the method described in section IIIB. We use both the per capita and the aggregate MSFE selection criteria given in table 2 with $\tau = 10$ periods ahead and $n = 10$ forecasts for each of the models. Thus, the earliest of our forecasts use all information up until 1982 to forecast carbon emissions for the year 1992, while the latest forecast used in the calculation of MSFE uses information up to 1991 to predict emissions in 2001. We then use all the information until 2001 to calculate future predicted emissions for each model to 2011.

In figure 2 we plot the historical CO_2 emissions from 1960 to 2001 and the density for the 5th to the 95th percentile of point forecasts for models in our universe until 2011. The top panel displays the distribution of point forecasts for the indirect models, which require forecasts of the right-hand-side variables. The bottom panel displays the distribution for the direct models. The two panels are different from a traditional predictive density, since they display the range of point forecasts from the universe of models instead of giving readers an understanding of the degree of confidence for each forecast. Each shade represents 5% of the point forecast distribution and displays the corresponding range of point forecasts across models. Figure 2 also shows the predictions of the two models, which minimize aggregate mean square forecast error (dashed line) and per capita mean squared forecast error (solid line). Since our objective is to forecast total U.S. carbon emissions, the preferred criterion is aggregate mean squared forecast error, as discussed in section IIIB. Therefore, any unqualified references in the remainder of this paper to the "best model" refer to the best model based on the aggregate MSFE selection criterion.

Figure 2 demonstrates how large the variation is in forecasts from the individual models in our universe. A substantial number of models predict extremely large increases in carbon emissions, while more than 6% of the models predict emissions in 2011 lower than those in 2001. This large variation in model performance is also suggested by the range of MSFE for the models in the universe, which ranges from 503 to 19 million. It is not surprising to have so many poorly performing models given the large number of possible specifications. In practice, some of these specifications may be judged to be inferior a priori by the modeler and therefore never be estimated. The strength of the approach in this paper is that we minimize the degree to which a priori judgment is used to decide which models to include in the model universe to be

[21] The RC test methodology is summarized in more detail in Appendix A.

FIGURE 2.—DENSITY OF POINT FORECASTS FOR ALL MODELS AND BEST OUT-OF-SAMPLE MODELS



The figure 2 shows historical CO_2 emissions from 1960 to 2001 and the density for the 5th to the 95th percentile of point forecasts for models in our universe until 2011. Each shade represents 5% of the distribution. The figure also shows the predictions of the two models, which minimize aggregate mean square forecast error (dashed line) and per capita mean squared forecast error (solid line).

searched over. The assumption that the true model is included in our universe is therefore more viable here than if we were to preselect models to be used for comparison.

Comparing the set of indirect and direct forecasts in their out-of-sample test performance reveals an interesting finding. The indirect forecasts dominate the direct forecasts in the MSFE rankings. There is no direct model among the top 100 based on per capita MSFE or based on the aggregate MSFE. This suggests that in our case, the extra information contained in the out-of-sample projections of the right-hand-side variables is valuable and improves forecast performance.

Figure 2 also shows that the model selected according to the per capita criterion predicts lower emissions for the year 2011 at 1,624 million tons compared to the best model, with 1,689 million tons of carbon. The difference of 65 million tons of carbon is equivalent to 4% lower emissions than the best model predicts or to 5% of 1990 U.S. carbon emissions.

The per capita model is less parsimonious compared to the best model. It includes a linear and a quadratic income term as well as lagged right-hand-side variables. Both models include a linear population variable. Both selected models are dynamic in nature, employing a lag structure of one period's carbon emissions, which is consistent with the nature of an installed durable capital stock. This is especially important in the carbon context, since carbon-emitting capital (as in power plants) is extremely durable.

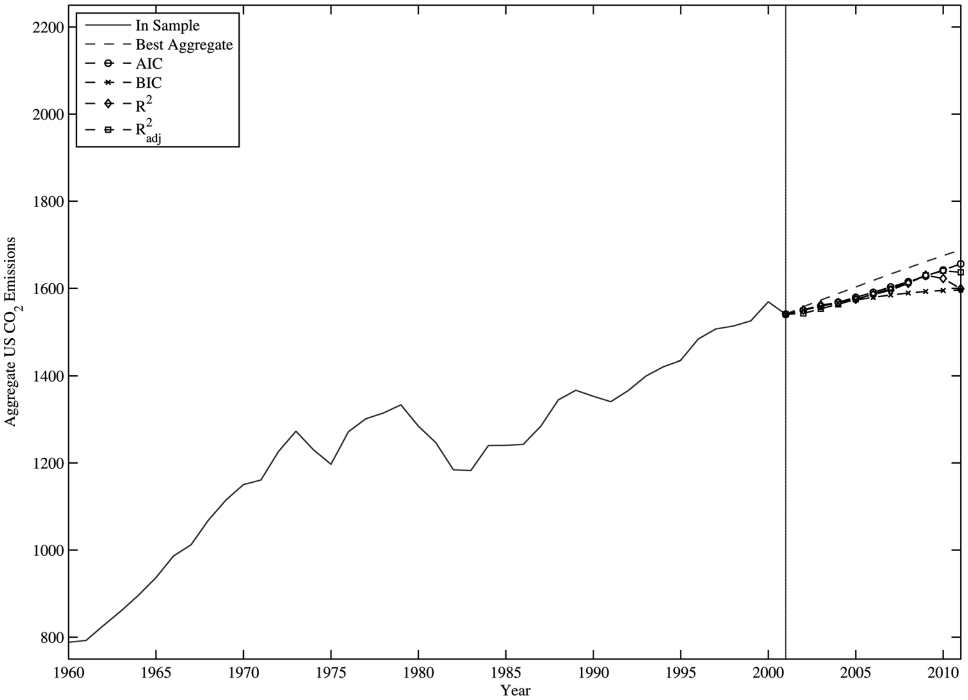
We formally test the superior predictive ability of the best model over the per capita MSFE model using the Hansen (2005) test. The RC test strongly rejects the null hypothesis that the best model has no predictive superiority over the benchmark per capita best model at the 1% level. This finding provides statistical evidence that using a per capita loss criterion when interested in the aggregate outcome of the series leads to selection of an inferior model specification.

B. *Performance of In-Sample Criteria*

Next, we compare the forecasting performance of models chosen by in-sample selection criteria with that of the best model. We choose the best model in sample based on four selection criteria: the Akaike information criterion (AIC), Schwarz information criterion (SIC), R^2, and $\bar{R}^2$.

Figure 3 shows the predicted emissions trajectories of the models selected by the in-sample selection criteria, with the best model forecast as comparison. The simple R^2 and adjusted R^2 selected models perform poorly in our prediction

FIGURE 3.—FORECASTS FROM MODELS SELECTED ACCORDING TO IN-SAMPLE CRITERIA



The figure 3 shows historical CO$_2$ emissions from 1960 to 2001 and the out-of-sample predictions of the models, which were selected based on the different in-sample criteria. The dashed line shows the predictions of the best aggregate model, which minimize aggregate mean square forecast error.

TABLE 3.—AGGREGATE AND PER CAPITA MSFE AND REALITY CHECK BOOTSTRAP TEST RESULTS FOR BEST IN-SAMPLE MODELS

Tested Benchmarks:	AIC	BIC	Adjusted R^2	Simple $\bar{R}^2$
MSFE aggregate CO$_2$	4,667	4,311	5,664	6273
MSFE per capita CO$_2$	306	277	361	379
Hansen's RC p-value	0.004	0.006	0.000	0.000

Aggregate and per capita mean square forecast errors for the best models according to the different in-sample selection criteria. The reality check bootstrap test (RC) tests the best model's predictive superiority using each of the four in-sample selected models in turn as a benchmark.

experiment, as seen in table 3. They predict 2011 emission levels 5% and 3% below those of the best model, according to the per capita MSFE measure.[22] The SIC and AIC criteria choose models that also predict a lower emissions path compared to the best model. They lie, respectively, 6% and 2% under the 2011 best model predictions.

As table 3 shows, we formally reject the null hypothesis that the best model fails to outperform the models selected based on the four in-sample fitting criteria. For all four models, the test rejects at the 1% level that the best MSFE model is not superior to each of them in its predictive ability to forecast aggregate U.S. carbon emissions. These results suggest that using in-sample selection criteria is likely to be suboptimal when the goal is to forecast aggregate emissions.

C. *Performance of Benchmark Models from the Literature*

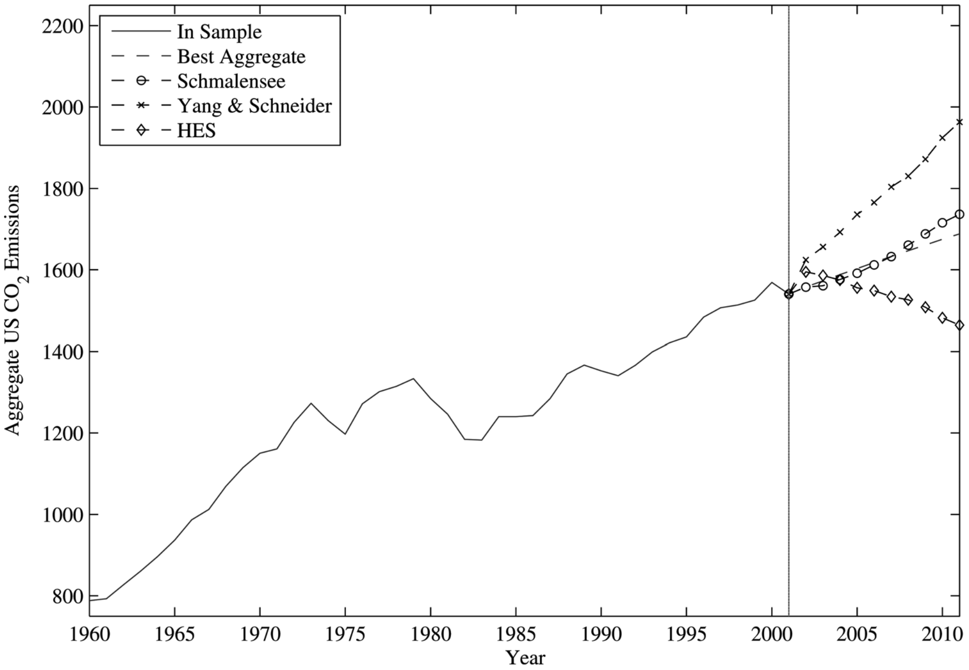
We now compare the performance of the best model with the performance of the three benchmark models found in the literature, which we discussed in section IIIC. Figure 4 shows the predictions from the three benchmarks and from the best model. We see that while the Holtz-Eakin & Selden (1995) model predicts a sharp downturn in carbon emissions, the Yang & Schneider (1998) model predicts a sharp increase of emissions. The Schmalensee et al. (1998) predictions lie closest to the best model among the three, with predicted emissions for the year 2011 of 1.74 billion tons of carbon.

When running the formal RC test, we again reject the null hypothesis that the best model has no superior predictive ability over the benchmark in all three cases. The results are presented in table 4, where we also present the MSFEs for each model. We can see why the null is strongly rejected for all three benchmark models, as their MSFEs from the out-of-sample experiment are much higher than that of the best-performing model. Furthermore, we find that the benchmark based on the model of Schmalensee et al. (1998) is outperformed by over 6,700, or about 25% of the models in our universe.

Also included in table 4 are the results of the RC test, with the best model itself as a benchmark. With the best model taken out as the benchmark, the best model encountered in the specification search is the model with the second-lowest aggregate MSFE in our model universe. That

[22] Both criteria select indirect forecast models with nonparsimonious specifications. Nonparsimony, especially for higher-order terms, indirectly punishes the models' forecasting performance, as each variable adds estimation uncertainty when variables need to be projected out-of-sample.

Figure 4.—Forecast Comparison of Benchmark Models from the Literature with the Best Model



Historical CO_2 emissions from 1960 to 2001 and the out-of-sample predictions of the benchmark models from the literature. The dashed line shows the predictions of the best model based on aggregate emissions.

Table 4.—Aggregate and Per Capita MSFE and Reality Check Bootstrap Test Results for Benchmark Models from the Literature

	Tested Benchmarks			
	Schmalensee, Stoker and Judson	Holtz-Eakin and Selden	Yang and Schneider	Best Model
MSFE aggregate CO_2	3,163	2,701	2,215	503
MSFE per capita CO_2	456	559	195	86
Hansen's RC p-value	0.000	0.000	0.000	0.940

Aggregate and per capita mean square forecast errors for well-known specifications from the reduced-form literature. The RC tests the best model's predictive superiority using each of the three models from the literature as the benchmark.

means we are effectively testing the null hypothesis that the second-best model in the model universe has no predictive superiority over the best model. The resulting p-value of 0.94 shows clearly that, as we would expect, we cannot reject the null in this case. Furthermore, we find in a RC test where we use the second-best model as the benchmark that the null hypothesis is not rejected; that is, the best and the second-best model are too close to statistically conclude that the best model has a better predictive ability. This indicates that we may have multiple models in the universe that are similar in their predictive ability to the best model.[23]

D. *Benefits from Disaggregation*

A large, emerging literature shows possible significant benefits from disaggregation of the data series used to forecast an aggregate (Marcellino et al., 2003; Auffhammer & Steinhauser, 2007). In order to check whether we get similar benefits from disaggregation, we run the following experiment. We use our best model specification based on the aggregate MSFE criterion and calculate its MSFE in such a way that it is comparable with a MSFE derived from an aggregate series. For each out-of-sample forecast, we first aggregate emissions across states to get aggregate U.S.-level CO_2 emissions instead of calculating the forecast error at the state level. We then subtract the model's forecast from the actual CO_2 emissions before the error is squared and repeat this experiment as before for ten time periods. This results in an MSFE of 1,414 for our best model.

As we did for the disaggregate model specifications, we now construct a ten-year-ahead forecasts for all feasible specifications using aggregate data on the national level.[24] The model with the smallest forecast error among these out-of-sample experiment forecasts has an MSFE of 3,651. At a forecast horizon of ten years out, we therefore get an improvement in MSFE of 61% for our best model when using disaggregate data compared to the best-performing model under aggregate U.S. data.

[23] We can run RC tests for all models in our universe, each individually as the test's benchmark, to see which models perform close to the best model. That way, we identified a set of nonrejectable models for which, when used as a benchmark, the Hansen p-value of the RC test is larger than 10%. The evenly weighted forecast average over all nonrejectable models based on the aggregate MSFE predicts 2011 emission of 1,608 million tons.

[24] This is possible only for a subset of 1,632 unique models that do not use state-specific variables or time fixed effects.

E. Comparison to other U.S. Carbon Forecasts

Using the Emission Scenario Database developed for the IPCC's *Special Report on Emissions Scenarios*, we can compare our predictions to scenario outcomes of previous studies under similar "no intervention" assumptions.[25] These mostly structural studies produce 71 forecasts for the United States under scenarios classified as no intervention.[26]

The different emission estimates of all studies for the United States for the year 2010 range from 1.30 to 2.36 billion tons of carbon, with a mean of 1.77. For the same year, our best model selected on aggregate emission loss predicts 1.67 billion tons.[27] Our forecast suggests emissions might be 100 million tons of carbon lower on average than what existing scenarios predict. This estimate is 5.6% less than the IPCC database average or a 7.4% difference with respect to 1990 U.S. emissions levels. A difference of such magnitude in projected business-as-usual emission would significantly overstate the costs of committing to a given emissions-reduction scheme.

The exercise above compares our forecasts to a rival set of forecasts for a future date. We cannot verify which actual does better. We conduct another exercise using forecasts made in the past and compare them to actual observed emissions data. For a subset of the studies in the IPCC Emission Scenarios Database (Morita, 1999), which were updated before the year 2000, we can use the forecast for the year 2000 and compare them with the forecast from our best model and the actual CO_2 emissions in that year. This is the case for 40 of the 71 scenarios. They were on average done in 1995 and predicted BAU emissions for the year 2000 of 1,591 million metric tons of carbon. Comparing this to the predictions of our best models based on 1990 information using a longer ten-year horizon, we predict 1,548 million tons of carbon emission for the United States in 2000. So for the year 2010, we find that the best-models predictions are lower than the average of the IPCC forecasts—in this case by 43 million tons, or 3%. The actual emissions in the year 2000 were 1,554 million tons of carbon. Our best model is about 0.4% off, while the IPCC database average, with an average horizon of five years, is about 2.4% off.

V. Conclusion

Reduced-form econometric models are an important alternative to more complex structural models for forecasting carbon emissions. While structural models are an essential tool for policy simulations, reduced-form models have some significant advantages if we are interested in nonintervention scenarios, for example, as input to global circulation models or as baseline to evaluate abatement commitments. Lower data requirements for reduced-form models allow for the use of longer time series and facilitate the analysis for countries where the structural approach is infeasible. However, the existing reduced-form literature on forecasting CO_2 emissions is characterized by a wide variety of conflicting specifications and resulting point forecasts. The primary source of this inconsistency among forecasts is the tendency to conduct a model search over a very limited set of the theoretically and empirically feasible models. Further, the confidence with which the superiority of a preferred model is claimed can be questioned due to a lack of accounting for model search in performance testing. This omission leaves open the strong possibility that identification of a best model was the result of a chance alignment of the model with the observed data rather than it being the best model of the underlying data-generating process.

The empirical application part of this paper provides forecasts of U.S. CO_2 emissions based on the best-performing model selected from a large universe of over 27,000 models. We use Hansen's (2005) reality check bootstrap test to formally compare models based on their predictive abilities, taking the data snooping or data mining of the model search into account. The test statistics show that the best model from our search significantly outperforms existing benchmark models found in the literature.

A more general contribution of this paper is to highlight the importance of the choice of model selection criterion. For most policy and climate modeling purposes, the output of interest is a forecast of aggregate carbon emissions. The existing literature has instead relied on in-sample ability to predict per capita emissions. We first demonstrate how the use of in-sample selection criteria leads to choosing inferior performing models when one is interested in the out-of-sample predictive ability. We select best models using four of the most popular in-sample performance criteria: R^2, $\bar{R}^2$, SIC, and the AIC. We then compare the performance of these four models against our model universe on the basis of out-of-sample performance. All four models selected based on in-sample performance are significantly outperformed by the best model in our universe.

We next show how the use of a loss function defined over per capita emissions rather than total emissions can lead to poor performance when the objective is to forecast total emissions. We find that the point forecasts of the best model selected on the basis of per capita loss are 4% lower by 2011 and diverge quickly from the trajectory predicted by the best model selected on the basis of minimizing a loss function defined over aggregate emissions. Furthermore, the reality check bootstrap Test confirms that the per capita model has significantly inferior predictive ability.

While much of the contribution of this paper is with respect to model selection, from a policy perspective, the contribution

[25] The database was assembled by Morita (1999) at the Center for Global Environmental Research at the National Institute for Environmental Studies and made available on its Web site under http://www.cger.nies.go.jp/scenario/index.html.

[26] Among these studies are IPCC IS92 scenarios, results from the Energy Modeling Forum, Nordhaus's RICE model, reports from the U.S. Energy Information Administration, and many others.

[27] Forty seven of the 71 models predict larger emissions, while 24 predict a lower carbon output for the United States. Interestingly, the EIA Energy Outlook from May 1996 predicts an almost identical level of 1.66 billion tons.

of interest is an improved reduced-form carbon forecast. Since forecasts for CO_2 emissions serve as important inputs to global climate models as well as inputs to benefit-cost studies, suboptimal forecasts may have real consequences for future global climate agreements. For example, nations with downward-biased projections of their business-as-usual emissions will underestimate the costs of committing to a given emissions-reduction scheme and might commit to overly stringent goals. Of course, the opposite is also possible, where countries may be reluctant to join an agreement based on upwardly biased projections of their BAU emissions, which will overstate the costs of committing to a given emissions reduction scheme. Our results suggest a 5.6% lower carbon dioxide emission prediction (or 7.4% in terms of 1990 U.S. emissions) than the average from the IPCC Emission Scenario Database for 2010. This is a large difference in context of the 7% of emission reductions relative to 1990 level the United States agreed to under the Kyoto Protocol.

REFERENCES

Aldy, Joseph E., "An Environmental Kuznets Curve Analysis of US State-Level Carbon Dioxide Emissions," *Journal of Environment and Development* 14:1 (2005), 48–72.

Arrow, Kenneth, Bert Bolin, Robert Constanza, Partha Dasgupta, Carl Folke, Crawford Stanley Holling, Bengt-Owe Jansson, Simon Levin, Karl-Göran Mäler, Charles Perrings, and David Pimentel 'Economic Growth, Carrying Capacity and the environment," *Science* 268 (1995), 520–521.

Auffhammer, Maximilian, "The Rationality of EIA Forecasts under Symmetric and Asymmetric Loss," *Resource and Energy Economics* 29 (2007), 102–121.

Auffhammer, Maximilian, and Richard T. Carson, "Forecasting the Path of China's CO_2 Emissions Using Province-Level Information," *Journal of Environmental Economics and Management* 55 (2008), 229–247.

Auffhammer, Maximilian, and Ralf Steinhauser, "The Future Trajectory of US CO_2 Emissions: The Role of State vs. Aggregate Information," *Journal of Regional Science* 47:1 (2007), 47–61.

Azomahou, Théophile, François Laisney, and Phu Nguyen Van, "Economic Development and CO_2 Emissions: A Nonparametric Panel Approach," *Journal of Public Economics* 90 (2006), 1347–1363.

Blasing, T. J., C. T. Broniak, and Gregg Marland, *Trends: A Compendium of Data on Global Change* (Oak Ridge, TN: Oak Ridge National Laboratory, U.S. Department of Energy, Carbon Dioxide Information Analysis Center, 2004).

Böhringer, Christoph, Klaus Conrad, and Andreas Löschel, "Carbon Taxes and Joint Implementation: An Applied General Equilibrium Analysis for Germany and India," *Environmental and Resource Economics* 24:1 (2003), 49–76.

Böhringer, Christoph, and Heinz Welsch, "Contraction and Convergence of Carbon Emissions: An Intertemporal Multi-Region CGE Analysis," *Journal of Policy Modeling* 26:1 (2004), 21–39.

Copeland, Brian R., and M. Scott Taylor, "Trade, Growth, and the Environment," *Journal of Economic Literature* 42 (2004), 7–71.

Cowles, Alfred, "Can Stock Market Forecasters Forecast?" *Econometrica* 1 (1933), 309–324.

Diebold, Francis X., and Roberto S. Mariano, "Comparing Predictive Accuracy," *Journal of Business and Economic Statistics* 13 (1995), 253–263.

Ehrlich, Paul R., and John P. Holdren, "Impact of Population Growth," *Science* 171:3977 (1971), 1212–1217.

Garbaccio, Richard F., Mun S. Ho, and Dale W. Jorgenson, "Controlling Carbon Emissions in China," *Environment and Development Economics* 4 (1999), 493–518.

Giacomini, Raffaella, and Clive W. J. Granger, "Aggregation of Space-Time Processes," *Journal of Econometrics* 118:1 (2004), 7–26.

Goldberger, Arthur S., "The Interpretation and Estimation of Cobb-Douglas Functions," *Econometrica* 36 (1968), 464–472.

Grossman, Gene M., and Alan B. Krueger, "Environmental impacts of a North American Free Trade Agreement," (pp. 13–57), in Peter M. Garber (Ed.), *The Mexico-U.S. Free Trade Agreement*, (Cambridge, MA: MIT Press, 1993).

Hansen, Peter R., "A Test for Superior Predictive Ability," *Journal of Business and Economic Statistics* 23 (2005), 365–380.

Harbaugh, Bill, Arik Levinson, and Dave Wilson, "Reexamining the Empirical Evidence for an Environmental Kuznets Curve," this REVIEW 84 (2002), 541–551.

Holtz-Eakin, D., and Thomas M. Selden, "Stoking the Fires? CO_2 Emissions and Economic Growth," *Journal of Public Economics* 57 (1995), 85–101.

Houthakker, Hendrik S., and Lester D. Taylor, *Consumer Demand in the United States*, 2nd ed. (Cambridge, MA: Harvard University Press, 1970).

IPCC, *Climate Change 1992: The Supplementary Report to the IPCC Scientific Assessment* (Cambridge, MA: Cambridge University Press, 1992).

——— *Climate Change 2001, Working Group II: Impacts, Adaptation and Vulnerability* (Cambridge, MA: Cambridge University Press, 2001a).

——— *Special Report on Emissions Scenarios* (Cambridge, MA: Cambridge University Press, 2001b).

——— *Climate Change 2007, Working Group I: The Physical Science Basis* (Cambridge, MA: Cambridge University Press, 2007).

Judson, Ruth A., and Ann L. Owen, "Estimating Dynamic Panel Data Models: A Guide for Macroeconomists," *Economics Letters* 65 (1999), 9–15.

Leamer, Edward E., *Specification Searches: Ad Hoc Inference with Nonexperimental Data* (New York: Wiley, 1978).

Lieb, Christoph M., "The Environmental Kuznets Curve and Flow versus Stock Pollution: The Neglect of Future Damages," *Environmental and Resource Economics* 29 (2004), 483–507.

Lo, Andrew W., and A. Craig MacKinley, "Data Snooping Biases in Tests of Financial Asset Pricing Models," *Review of Financial Studies* 3 (1990), 431–468.

Marcellino, Massimiliano, James H. Stock, and Mark W. Watson, "Macroeconomic Forecasting in the Euro Area: Country Specific versus Area-Wide information," *European Economic Review* 47:1 (2003), 1–18.

Marland, Gregg, T. A. Boden, and R. J. Andres, *Trends: A Compendium of Data on Global Change* (Oak Ridge TN: Oak Ridge, National Laboratory, U.S. Department of Energy, Carbon Dioxide Information Analysis Center, 2004).

McCracken, Michael W., and Kenneth D. West, "Inference about Predictive Ability," (pp. 299–321), in Michael P. Clements and Devid F. Hendry (Eds.), *A Companion to Economic Forecasting*, (Malden, MA: Blackwell, 2004).

Millimet, Daniel, John A. List, and Thanasis Stengos, "The Environmental Kuznets Curve: Real Progress or Misspecified Models?" this REVIEW 85 (2003), 1038–1047.

Morita, Tsuneyuki, "Emission Scenario Database." For IPCC Special Report on Emission Scenarios (Geneva: IPCC, 1999).

Nickell, Stephen J., "Biases in Dynamic Models with Fixed Effects," *Econometrica* 49 (1981), 1417–1426.

O'Neill, Brian C., and Mausami Desai, "Accuracy of Past Projections of US Energy Consumption," *Energy Policy* 33 (2005), 979–993.

Politis, Dimitris N., and Joseph P. Romano, "The Stationary Bootstrap," *Journal of the American Statistical Association* 40 (1994), 1303–1313.

Schmalensee, Richard, Thomas M. Stoker, and Ruth A. Judson, "World Carbon Dioxide Emisions: 1950–2050," this REVIEW 80 (1998), 15–27.

Selden, Thomas M., and Daqing Song, "Environmental Quality and Development: Is There a Kuznets Curve for Air Pollution Emissions," *Journal of Environmental Economics and Management* 27 (1994), 147–162.

Stock, James H., and Mark W. Watson, "Macroeconomic Forecasting Using Diffusion Indexes," *Journal of Business and Economic Statistics* 20 (2002), 147–162.

Sullivan, Ryan, Allan Timmermann, and Halbert White, "Data-Snooping, Technical Trading Rule Performance, and the Bootstrap," *Journal of Finance* 5 (1999), 1647–1691.

U.S. Energy Information Administration, "Annual Energy Outlook 2004 with Projections to 2025," DOE/EIA report no. 0383 (2004).

Vollebergh, Herman R. J., Bertrand Melenberg, and Elbert Dijkgraaf, "Identifying Reduced-Form Relations with Panel Data: The Case of Pollution and Income," *Journal of Environmental Economics and Management* 58 (2009), 27–42.

Wallis, Kenneth, "Macroeconomic Forecasting: A Survey," *Economic Journal* 99 (1989), 28–61.

West, Kenneth D., "Asymptotic Inference about Predictive Ability," *Econometrica* 64 (1996), 1067–1084.

Wetrogan, Signe I., *Provisional Projections of the Population of States, by Age and Sex, 1980 to 2000*, (Washington, D.C.: U.S. Department of Commerce, Bureau of the Census, 1983).

White, Halbert, "A Reality Check for Data Snooping," *Econometrica* 68 (2000), 1079–1126.

Yang, Christopher, and Stephen H. Schneider, "Global Carbon Dioxide Emissions Scenarios: Sensitivity to Social and Technological Factors in Three Regions," *Mitigation and Adaptation Strategies for Global Change* 2 (1998), 373–404.

APPENDIX A

The Reality Check Bootstrap Test

First developed by White (2000) and later modified by Hansen (2005), the reality check bootstrap test allows undertaking data snooping or data mining with a degree of confidence that one will not mistake results generated by chance for genuinely good results.[28] The test gives a measure of confidence that the encountered model's predictive ability is not just a fluke of our model search. The null hypothesis, that the best model encountered during the specification search has no predictive superiority over a given benchmark model, takes the form

$$H_0 : \max_{k=1,\dots,L} E[f_k] \leq 0,$$

where $E[f_k] = \bar{f}_k = n^{-1}\sum_t \hat{f}_{k,t+\tau}$ and $\hat{f}_{k,t+\tau} = \sum_{i=1}^{50} -(c_{i,k,t+\tau} - \hat{X}_{i,k,t+\tau}\hat{\beta}_{i,k,t})^2 + (c_{i,0,t+\tau} - \hat{X}_{i,0,t+\tau}\hat{\beta}_{i,0,t})^2$. $\hat{X}_{i,k,t+\tau}$ includes all projected right-hand-side variables of a model k at the out-of-sample period $t + \tau$ for state i, $\hat{\beta}_{i,k,t}$ is an estimate that incorporates all information up to period t, and $c_{i,t+\tau}$ is the actual realization of the dependent variable—carbon emissions, in our case. The model $k = 0$ is the benchmark model so that under the null hypothesis, we expect the benchmark to outperform the best of all models contained in the universe. The alternative is that the best model is superior to the benchmark. This formulation of $\hat{f}_{k,t+\tau}$ is based on an MSFE selection criterion, but other criteria could be chosen. Hansen (2005) suggests a studentized test statistic and shows its improved power properties compared to White's former statistic. The Hansen test statistic takes the form

$$T^{RC} = \max_{k=1,\dots,L} \frac{n^{1/2}\bar{f}_k}{\sqrt{\text{var}(n^{1/2}\bar{f}_k)}}.$$

The difficulty in finding the distribution for T^{RC} is overcome by a bootstrap implementation. The bootstrap resamples the $\hat{f}_{k,t+\tau}$ to construct a distribution for T^{RC} and obtain the p-value of the test statistic. We used the stationary bootstrap of Politis and Romano (1994) with a block length of 4 and $B = 500$ resamples. The resampled statistic is in accordance with MSFE criteria computed as

$$\bar{f}_{k,b} = n^{-1}\sum_t \hat{f}_{k,\theta_{b,t}+\tau} \qquad \forall b = 1,\dots,B,$$

where, in our case, $n = 10$ and $t = 1982,\dots,1991$. We seek the distribution of the test statistics under the null hypothesis, so we impose the null by recentering the bootstrap variables. White (2000) proposes centering about $\bar{f}_k$ and shows that the distribution of $\bar{f}_{k^*} = \max_{k=1,\dots,L}\bar{f}_k$ is properly approximated by

$$T^{RC*}_{u,b} = \max_{k=1,\dots,L} n^{1/2}(\bar{f}_{k,b} - \bar{f}_k) \qquad \forall b = 1,\dots,B.$$

Hansen demonstrates that this most conservative approach gives too much power to poorly performing models with high variances in $\bar{f}_k$. The poorly performing models can dominate the right-hand tail of the distribution when a large, negative $\bar{f}_k$ is subtracted from a much less negative resampled $\bar{f}_{k,b}$. We find this to be a problem for our large range of models and their varying performances in our model universe. Poor models shift the distribution of the test statistic so far to the right that even for a perfectly predicting model, we could not reject the null. Hansen (2005) proposes that failure to center the poorly performing high-variance models around the mean presents an alternative distribution for $\bar{f}_{k^*}$:

$$T^{RC*}_{c,b} = \max_{k=1,\dots,L} n^{1/2}\frac{\bar{f}_{k,b} - \bar{f}_k \cdot \mathbf{1}_{\{\bar{f}_k \geq -A_k\}}}{\widehat{var}(n^{1/2}\bar{f}_k)} \qquad \forall b = 1,\dots,B,$$

where $A_k = 1/4n^{-1/4}\sqrt{\widehat{\text{var}}(n^{1/2}\bar{f}_k)}$ and $\widehat{\text{var}}(n^{1/2}\bar{f}_k) = B^{-1}\sum_b(n^{1/2}\bar{f}_{k,b} - n^{1/2}\bar{f}_k)^2$. To calculate the RC p-value, we sort the values $T^{RC*}_{c,b}$, denote them as an order statistic $T^{RC*}_{c,1}, T^{RC*}_{c,2},\dots,T^{RC*}_{c,B}$, and find the N for which $T^{RC*}_{c,N} \leq T^{RC} < T^{RC*}_{c,N+1}$. The reality check bootstrap p-value is now defined as

$$p_c = 1 - N/B.$$

Alternatively one could fit a suitable density model to the order statistic and get the p-value from the fitted distribution.

APPENDIX B

Model Specifications

Here we spell out in detail the model specifications of selected models described in sections IVA and IVB. Following the same notation as used in equation (1) and throughout the paper, the specification of the best model based on the aggregate MSFE is

$$\ln(c_{i,t}) = \alpha_0 + \rho_i \ln(c_{i,t-1}) + \alpha_1 \ln(\text{pdens}_{i,t}) + \alpha_2\text{oil}_i + \varepsilon_{i,t}, \tag{4}$$

where $\ln(c_{i,t})$ are log per capita carbon emissions for state i in year t, pdens$_{i,t}$ is population density, α_0 is the coefficient of a constant term, oil$_i$ is one for oil- or gas-producing states i, and $\varepsilon_{i,t}$ is assumed to be a stationary ergodic error term.

The model with the lowest per capita MSFE has the following specification:

$$\begin{aligned}\ln(c_{i,t}) = {}& \alpha_0 + \rho_i \ln(c_{i,t-1}) + \alpha_1 \ln(incomepc_{i,t}) + \alpha_2(\ln(incomepc_{i,t}))^2 \\ & + \alpha_3 \ln(pdens_{i,t}) + \alpha_4 \ln(incomepc_{i,t-1}) \\ & + \alpha_5(\ln(incomepc_{i,t-1}))^2 + \alpha_6 \ln(pdens_{i,t-1}) \\ & + \alpha_7 crisis_t + \alpha_8 coal_i + \varepsilon_{i,t},\end{aligned}$$

with the variables as defined above and where $\ln(incomepc_{i,t})$ is log per capita real personal income for state i in year t; $crisis_t$ is a vector of three time-varying dummies that switch to 1 for the years 1973–1975, 1979–1981, or 1990–1991, respectively; and coal$_i$ is 1 for coal-producing states. The other variables are defined as in equation (4).

Now we turn to the models selected based on in-sample information criteria. The model performing best according to the Schwarz information criterion has the following specification:

$$\begin{aligned}\ln(c_{i,t}) = {}& \alpha_0 + \rho_i \ln(c_{i,t-1}) + \alpha_1 \ln(incomepc_{i,t}) + \alpha_2(\ln(incomepc_{i,t}))^2 \\ & + \alpha_3 coastal_i + \gamma_t + \varepsilon_{i,t},\end{aligned}$$

where $coastal_i$ is a dummy variable that is 1 for coastal states and γ_t is a year fixed effect.

The model with the lowest Akaike information criterion value has the following specification,

$$\begin{aligned}\ln(c_{i,t}) = {}& \alpha_0 + \rho_i \ln(c_{i,t-1}) + F^6(\ln(incomepc_{i,t})) + \alpha_1 coastal_i \\ & + \alpha_2\text{oil}_i + \gamma_t + \varepsilon_{i,t},\end{aligned}$$

where $F^6(\cdot)$ represents a spline (piecewise linear) function with six segments and the other variables are defined as above.

[28] Hansen (2005) refers to his variation of the test as superior predictive ability (SPA) test. We will keep using the name reality check (RC) throughout.