Allocating emissions permits in cap-and-trade programs: Theory and evidence *

Preliminary.

Please do not cite without permission

Comments very welcome

March 2010

Meredith Fowlie†

Abstract

The allocation of emissions permits in "cap-and-trade" programs is an increasingly contentious policy design issue. Recent theoretical work has characterized the efficiency and distributional implications of alternative approaches to allocating these permits in detail. This paper tests whether observed firm behavior is consistent with the standard theory. I develop a simple analytical model to capture the essential theoretical relationships between permit allocation design choices and short-run production decisions. Data gathered from a multi-state emissions trading program are then used to analyze these relationships empirically. Results suggest that larger and more polluting firms incorporate both the explicit environmental compliance costs (i.e. the costs of holding permits to offset emissions) and the less salient production subsidies implicitly conferred by dynamic permit allocation updating into their operating decisions. Among smaller and/or cleaner producers, I cannot reject the null hypothesis of a zero average effect.

* I thank Stephen Holland, Erin Mansur, and seminar participants at Environmental Defense, Harvard's Kennedy School, the University of California, Berkeley, the University of Michigan, the University of Toronto, and Yale University for helpful comments and discussion. I gratefully acknowledge support from the National Science Foundation through grant SES-0922401, the UC Energy Institute and the Center for State, Local, and Urban Policy at the University of Michigan. Erik Johnson provided excellent research assistance.

†Department of Agricultural and Resource Economics, University of California Berkeley, 301 Giannini Hall, Berkeley CA, 94720. email: fowlie@berkeley.edu

1 Introduction

Billions of dollars worth of tradable emissions permits are allocated each year to U.S. industrial producers regulated under emissions "cap-and-trade" programs.[1] In theory, how these permits are allocated can have significant implications for who will bear the costs and how efficiently the mandated emissions reductions will be achieved. Permit allocation has thus emerged as one of the more contentious issues in permit market design.

Regulatory agencies have been allocating tradable emissions permits under the auspices of local, regional, and nationwide cap-and-trade programs for over a decade. Over this time period, theoretical analyses of the efficiency and distributional implications of permit market design choices have grown increasingly sophisticated. However, we know relatively little about how permit allocation design affects firm decision-making in real world settings. This paper brings evidence to bear on a first-order empirical question: are firms responding to permit allocation incentives as standard theory predicts?

Traditionally, policy makers have chosen between two general approaches to allocating emissions permits: auctioning and grandfathering. Under an auction regime, emissions permits are sold to the highest bidder. In contrast, "grandfathered" permits are freely distributed to regulated sources based on pre-determined, firm-specific characteristics. In the absence of other market failures, this choice between grandfathering and auctioning should have no bearing on permit market efficiency in the short-run (Montgomery, 1974).

Many economists favor auctioning on the grounds that revenues can be used to offset distortionary taxes (Crampton and Kerr, 2002; Goulder *et al.*,1999).[2] However, in practice, policy makers have routinely chosen to forego auction revenues in favor of handing permits out for free to regulated entities.[3] The ability to make concessions to adversely impacted and politically pow-

[1] The value of permits allocated to emitting facilities under the NOx Budget Program and the Acid Rain Program each year is roughly \$1.4 B and \$4.5B, respectively. The U.S. Environmental Protection Agency (EPA) has estimated that the value of allowances allocated annually under proposed Federal climate legislation would exceed \$200 billion (US EPA, 2008) .

[2] Other efficiency-related arguments in favor of auctioning pertain to the mitigation of pre-existing regulatory distortions and distributional concerns. For example, Dinan and Rogers (2002) and Parry (2004)**?** emphasize the potential distributional implications of the allocation design choice, demonstrating that high income individuals are likely to gain more from freely allocated allowances than are low income individuals,

[3] A majority of permits are distributed freely to regulated entities in Southern California's RECLAIM program,

erful stakeholders via grandfathering has perhaps been as important a factor in the widespread adoption of emissions trading programs as the promise of cost minimization and gains from trade.

More recently, a third design alternative has emerged. Under a "contingent allocation" regime, updating rules established ex ante are used to determine how a firm's permit allocations will be periodically updated over the course of the trading program. Allocation updating is typically based on a firm's production choices (such as output levels or fuel inputs). The incentives created by contingent allocation rules are quite different from those associated with grandfathering or auctioning because updating creates an incentive to increase whatever activity determines future emissions permit allocations.

In a theoretical, "first-best" setting, it is straightforward to demonstrate that periodically updating firms' future permit allocations based on present production choices will undermine the efficiency of permit market outcomes because the implicit subsidy conferred by allocation updating encourages firms to increase output to economically inefficient levels. (Bohringer and Lange, 2005; Sterner and Muller, 2008).[4] However, contingent updating can welfare dominate more standard permit allocation approaches when there are additional, pre-existing distortions to contend with. For example, the theory literature has explored how allocation updating can be used reduce inefficiencies resulting from the exercise of market power (Fischer, 2003; Neuhoff, Martinez, and Sato, 2006), tax interaction effects (Fischer and Fox, 2007), and emissions leakage (Bernard *et al.,* 2007; Quirion and Demailly, 2006). Allocation updating can also be used to alter the distribution of costs of complying with a cap-and-trade program across producers and consumers (Jensen and Rasmussen, 2000).

Political support for contingent allocation updating is increasing. Industry groups endorse it as a "common sense way to promote efficiency, fairness, and environmental protection".[5] Policy experts concede that it may offer the most pragmatic approach to mitigating the adverse effects of environmental regulation on domestic industry competitiveness when emissions regulations are

---

the European Union's Emissions Trading Program, the nation-wide Acid Rain Program, and the regional NOx Budget Trading Program.

[4]Here, "first-best" refers to a regulatory environment in which the only market distortion or imperfection is the environmental externality that the emissions regulation is designed to internalize.

[5]"Advantages of Allocating Emissions Credits Based on Efficiency." The Clean Energy Group. May 15, 2009. http://www.thecleanenergygroup.com/lsgbriefings.asp

incomplete.[6] Federal climate legislation passed in June 2009 by the House of Representatives includes provisions for allocation updating as a means of compensating trade exposed emitters for compliance costs incurred.[7] In California, legislators are considering allocation updating as a means of mitigating impacts on consumer prices and reducing emissions leakage to unregulated entities (CPUC-CEC, 2008).[8]

Can permit allocation design be effectively used to achieve these kinds of policy objectives? This depends in large part on whether firms respond to permit market incentives as standard theory predicts. In policy debates, stakeholders have questioned the extent to which the implicit subsidy conferred by updating will be factored into firms' "real world" production decisions.[9] Others contend that the effects of the subsidies conferred by allocation updating will be mitigated by pre-existing regulatory and market constraints (NCEP, 2008). The academic literature offers a range of alternative models of private sector decision making that would lead firms to discount- or even ignore- the implicit subsidy conferred by allocation updating. For example, researchers have documented gain-loss asymmetries, asymmetric pass-through of operating costs, and heuristic approaches to dealing with cognitive constraints in different private sector contexts (see, for example, Duxbury and Summers, 2004; Hirshliefer, 2001 ; Zachmann and von Hirschhausen, 2008). If these institutional factors and/or behavioral phenomena affect how firms percieve and respond to emissions permit market incentives, permit allocation design features may not have the intended effect.

This study makes three contributions to both the academic literature and the ongoing policy discourse. First, a partial equilibrium model is used to illustrate the essential short-run implications of the permit allocation design. In an effort to clarify the terms of the policy debate,

[6] See, for example, the testimony of Richard Morgenstern of Resources for the Future. *Competitiveness and Climate Policy: Avoiding Leakage of Jobs and Emissions*: *Hearings before the Committee on Energy and Commerce U.S. House of Representatives.* March 18, 2009

[7] The justification for contingent updating in this context rests on the concern that firms would divert new investments and production to manufacturing facilities located in countries without commensurate regulations. Contingent allocation updating, intended as a stop-gap measure, compensates firms for the compliance costs incurred so as to mitigate adverse competitiveness impacts.

[8] The design recommendations of both the California Public Utilities Commission and the WCI include the minimization of the impacts of carbon regulations on consumers and the mitigation of leakage as objectives of the allocation process.

[9] See, for example, RGGI, 2004.

these implications are sometimes over-simplified to the point of misrepresentation. For example, a recent and influential report that seeks to "clear up misperceptions, common among many stakeholders, about how allocation decisions do and do not affect the way an emissions trading program works in practice" asserts that "[a]location affects the distribution of benefits and burdens among firms and industry sectors; it does not change program results or overall costs" (NCEP, 2007). In fact, this conventional wisdom does not hold when future permit allocations are contingent upon current production choices. The model intuitively demonstrates how permit allocation design can significantly affect firms' short-run production choices and thus aggregate social costs.

Second, the study provides insight into how variation in permit market prices is affecting firms' short-run production decisions. An emerging literature examines the relationships between permit price series and wholesale electricity price series (see, for example, Bunn and Fezzi, 2007; Fell, 2009; Sijm et al., 2006). Given the complexity of interactions between permit markets and wholesale electricity markets, it has been difficult to deduce firm-level behaviors from wholesale market price dynamics. Rather than try to evaluate the effect of permit market incentives on market-level outcomes, this paper examines firm-level responses to variation in permit market incentives in unprecedented detail.

Finally, the paper offers some of the first empirical evidence of how contingent allocation updating is working in practice. Previous attempts to study the effects of contingent permit allocation updating, and revenue recycling more generally, have been unable to convincingly separate the effects of the implicit subsidy from the overall effect of the environmental regulation.[10] I exploit an unusual policy setting in which the implicit subsidies conferred by a regional emissions trading program vary systematically across producers and across seasons. Much of this variation is exogenous- in an econometric sense- to firms' short-run production decisions.

The empirical analysis begins by demonstrating that statistical relationships in the data are

[10] Sterner and Isaksson (2006) **?** were the first to empirically investigate the effects of revenue recycling in the context of market-based emissions regulation. They analyze a Swedish program in which emissions charges are refunded to polluting firms in proportion to output. Because rebates do not vary across firms or across time, the authors cannot separate the effect of the tax fro the effect of the recycled revenues. Sjim *et al.* (2006) look at how sequentially grandfathered permits in the EU ETS impact electricity market outcomes. The lack of clarity surrounding how current production decisions will influence future permit allocations in the European Union's Emissions Trading System complicates their analysis of how this implicit updating has affected firm decision making.

generally consistent with standard theoretical predictions regarding producers' response to permit allocation incentives. Derivation of an estimable econometric model begins with a deterministic representation of the multi-period unit commitment problem that plant managers are presumably solving as they make their short run supply decisions. Focusing on one dimension of this dynamic problem - namely the decision to resume production conditional on being inactive- simplifies the derivation of a tractable reduced form that can be implemented empirically using a discrete choice framework. Conditional on the assumptions of the model, this provides a basis for testing hypotheses about the causal relationships between permit allocation design features and firms' short-run production decisions. Overall, I fail to reject the null hypothesis that electricity supply decisions are, on average, unaffected by variation in NOx permit prices. However, among larger producers, the supply response to changing permit prices is more consistent with the standard theory. The effect of a given change in the permit price is more negative among producers with relatively high emissions rates and less negative among producers receiving implicit production subsidies. I fail to reject the hypothesis that these producers are, on average, equally attentive to the pollution disincentive and production incentives conferred by allocation updating.

The paper proceeds as follows. Section 2 develops a simple theoretical model that is used to intuitively demonstrate the first order implications of alternative permit allocation designs. Section 3 introduces the NOx Budget Program. Section 4 summarizes the data. Section 5 discusses the underlying data generating process and lays out an empirical strategy. Section 6 presents the estimation results. Section 7 concludes.

2 Allocating emissions permits: Theory

The analytical framework introduced in this section serves two purposes. The model is first used to demonstrate the most essential short run implications of the permit allocation design choice.[11] The model then serves as a jumping off point for the empirical analysis.

[11] The analysis will focus on short-run relationships exclusively. To the extent that allocation updating is seen as a way to smooth the transition to auctioning regimes, these short-run relationships will be very important. Also, a clear characterization of short-run interactions is an essential first step towards understanding longer-run outcomes and implications.

The model is intentionally simple. Many of the institutional details and market imperfections captured by models found elsewhere in the literature (such as pre-existing tax distortions, the exercise of market power, or incomplete regulation) have been stripped away. Eliminating some of the complexities of real policy settings helps to highlight the most basic trade-offs between static production efficiency, static allocative efficiency, and distributional concerns.

2.1 A static, partial equilibrium framework

Production of a homogeneous good generates harmful pollution. Industry production in time t is denoted Q_t. Let q_{it} denote the quantity produced by firm i in time t. Producers are characterized by increasing marginal cost technologies; $C_i(q_i)$ and $c_i q_{it}$ denotes the unit-specific total cost function and marginal cost, respectively. Efficient factor markets are assumed. Marginal operating costs thus reflect the true opportunity cost of allocating inputs to production in this industry. Emissions rates e_i are constant per unit of output. Demand is characterized by an affine inverse demand function $P_t = a - bQ_t$.

To keep the model transparent and tractable, only two price taking firms are represented. A more general model with $N > 2$ is easily formulated but more difficult to intuitively interpret. The two firms are indexed c (denoting the relatively "clean" producer) and d (denoting the relatively "dirty" producer). For the purpose of this example, I assume emissions rates are negatively correlated with operating costs: ; $e_c < e_d$, $c_c > c_d$.[12]

Industry emissions are regulated under a cap-and-trade program; aggregate emissions in period t cannot exceed an exogenously determined cap E_t. The time path of permitted emissions ($i.e.$ $E_1, E_2, ...$) is set by the regulator ex ante. To comply with the program, firms must offset uncontrolled emissions with permits. These permits are tradable in an emissions permit market. There are no spatial or temporal restrictions on permit trading. I assume that firms acts as price takers in both the permit and product markets.

This short-run analysis conditions on existing production technology and operating charac-

[12] This assumption finds empirical support in the emissions market I consider here. Unit-level fuel operating costs and NOx emissions rates are negatively correlated in the data analyzed in the subsequent section. However, there are certainly examples of low-emitting facilities with relatively low operating costs, and high-emitting facilities with relatively high operating costs.

teristics; emissions rates and operating costs are exogenously determined and fixed. Emissions reductions can thus be achieved in two ways: through increasing the share of the market served by the relatively clean producer or reducing the quantity consumed.

In existing and planned cap-and-trade programs, permits are allocated via auctioning, grandfathering, symmetric "output-based" updating, or asymmetric updating. Stylized representations of these different approaches are considered below.

2.2 The benchmark case

Outcomes under alternative allocation rules will be compared against a "first best" benchmark that maximizes total economic surplus $S(Q)$ subject to technology operating constraints and the constraint that aggregate emissions do not exceed the exogenously set cap:

$$\begin{aligned} \max_{q_{ct}, q_{dt}} \quad &: \quad S(Q_t) = \int_0^{Q_t} P(Q_t) dQ_t - C_c(q_{ct}) \; - C_d(q_{dt}) \qquad (1) \\ s.t. \;\; e_c q_{ct} + e_d q_{dt} \quad &= \quad E_t \\ q_{ct} + q_{dt} \quad &= \quad Q_t. \end{aligned}$$

Note that the welfare measure $S(Q_t)$ reflects the utility associated with total consumption less production costs but does not capture the benefits associated with emissions reductions. Because aggregate emissions are held constant at E across all scenarios, changes in $S(Q_t)$ will reflect changes in absolute welfare vis a vis this benchmark.

The first order conditions for this maximization problem imply:

$$P(Q_t^*) - c_i q_{it}^* - \tau_t e_i = 0, \quad i = c, d. \qquad (2)$$

where τ_t is the shadow value of the emissions constraint at time t. The * superscript denotes values that maximize economic surplus subject to the constraint.

Rearranging these first order conditions (and omitting the t subscripts for expositional clarity) yields:

$$\frac{P - c_d q_d^*}{P - c_c q_c^*} = \frac{e_d}{e_c}. \tag{3}$$

Figure 1 helps to illustrate this result. The downward sloping line, representing the emissions constraint, connects all allocations of production across the two firms that exactly satisfy the emissions cap. The slope of this line is $\frac{e_d}{e_c}$. The economic surplus function $S(Q)$ is also projected into this space. The level sets of the surplus function appear as concentric iso-surplus curves. The slope of an iso-surplus curve measures the rate at which production at the clean firm can be substituted for production at the dirty firm while holding total economic surplus constant. The socially optimal allocation of production occurs at the point where the emissions constraint is just tangent to an iso-surplus curve. All other points that exactly satisfy the emissions constraint are associated with lower levels of economic surplus. The broken line connects all points associated with an aggregate production level Q^*. Production allocations on the emissions constraint lying strictly above (below) the optimal outcome are associated with more (less) consumption than is consistent with compliance constrained economic surplus maximization.

Two efficiency properties of this equilibrium are worth highlighting:

Property 1 : Marginal abatement costs (measured in terms of foregone profits per unit of emissions reduction) are set equal across producers:

$$\frac{P - c_d q_d^*}{e_d} = \frac{P - c_c q_c^*}{e_c}. \tag{4}$$

This assures that abatement activities have been efficiently allocated among producers. Given production level Q^* , the cost of meeting the emissions constraint E is minimized.

Property 2 : Emissions abatement activities are allocated efficiently across the supply and demand-side of the product market:

$$\frac{\partial S^*}{\partial E} = \frac{c_c q_c^* - c_d q_d^*}{e_d - e_c}. \tag{5}$$

Intuitively, the derivative of the welfare function with respect to the emissions constraint captures the marginal cost of reducing emissions via conservation measures on the demand side (i.e. through a reduction in consumption). The marginal abatement cost on the supply side is the cost of reallocating production from the low cost, high emitting producer to the high cost, low emitting producer so as to incrementally reduce emissions. Equation [5] implies that an optimal balance is struck between the two short-run abatement options. This result is derived in Appendix 1.

Taken together, these two efficiency properties imply that this equilibrium outcome minimizes the total economic cost of achieving the mandated emissions reductions.

2.3 Grandfathering and auctioning regimes

I now consider a perfectly competitive industry subject to a market-based emissions cap-and-trade program. Let A_{it} represent the permit allocation to firm i in period t. Under grandfathering, the number of permits the firm receives (free of charge) from the regulator each period is determined at the outset of the program. Under auctioning, $A_{it} = 0 \; \forall \; i$. Under either scenario, firms' future permit allocations are independent of their production decisions going forward.

Let τ_t represent the permit price (an endogenously determined parameter). The cost of holding permits to offset uncontrolled emissions is $\tau_t e_i q_{it}$. The profit maximization problem faced by price taking firm i in time period t is thus:

$$\max: \; \pi_{it} = P_t q_{it} - C(q_{it}) + \tau_t (A_{it} - e_i q_{it}), \qquad i = c, d. \tag{6}$$

Assuming price-taking behavior in both the permit and product markets, the first order conditions for this profit maximization problem are given by [2]. Thus, the efficiency properties [4] and [5] are achieved under both grandfathering and auctioning.

2.4 Contingent allocation updating

In a contingent updating regime, the total quantity of emissions permits to be allocated in each period, and the rules specifying how firms' production decisions will determine future permit allocations, are determined at the outset of the program. For example, consider the simplest of output-based allocation rules wherein a firm's permit allocation in period $t + 1$ is determined by its product market share in period t:

$$A_{it+1} = \frac{E_{t+1}}{Q_t} q_{it} \equiv s_{it} q_{it}. \tag{7}$$

The subsidy s_{it} is measured in terms of permits allocated to unit i in period $t+1$ per unit of output in period t. The size of this subsidy will depend on the total number of permits allocated in the future period E_{t+1}, how the firm discounts future revenue streams $\delta_i(t)$, the future permit price τ_{t+1}, and total industry production Q_t.

The allocation rule summarized by [7] is "symmetric" in the sense that the implicit subsidy parameter s_{it} does not vary across firms. In practice, contingent updating is often asymmetric. For political and practical reasons, the implicit subsidy per unit of output can vary across units with different technology types, fuel efficiency, or other observable operating characteristics.[13]

In general, contingent allocation updating adds an additional argument to the firm's profit function:

$$\max: \ \pi_{it} = P_{it} q_{it} - C(q_{it}) - \tau_t e_i q_{it} + \delta_i(1) \ \tau_{t+1} s_{it} q_{it}. \tag{8}$$

Here I assume that the updating parameters are defined such that the total number of permits allocated through updating does not exceed the total cap. This implies that the average updating parameter cannot exceed the average emissions rate. I also assume that updating occurs with a one period lag.

Some additional assumptions further simplify the analysis. Unrestricted banking and borrowing of permits, rational expectations, and zero arbitrage together imply that permit prices are constant in present value terms. If all firms discount future period gains at the market rate and take total sector output Q_t as given, the implicit subsidy per unit of production in period t simplifies to τs_t.[14] Omitting t subscripts for simplicity, the first order conditions for profit maximization in an allocation updating regime imply:

[13]For instance, many states in the NOx Budget Program use heat input based updating rules. Under these rules, the per-unit subsidy is greater for less fuel efficient firms.

[14]Alternatively, firms could take into account how their own production decisions affect aggregate production levels (and thus the size of the implicit subsidy). This would reduce the perceived production subsidy by $\frac{\delta_t \tau_t A_{t+1} q_t}{Q_t^2}$. Intuitively, if the firm incrementally increases production in time t, it decreases the number of permits allocated in time $t+1$ per share of output Q_t, although it is now entitled to an additional share.

$$\frac{P - c_d q_d'}{P - c_c q_c'} = \frac{e_d - s_d}{e_c - s_c},$$

where the superscript ' denotes equilibrium outcomes under contingent permit updating. The equilibrium permit price under contingent updating is given by :

$$\tau' = \frac{c_c q_c - c_d q_d}{e_d - e_c + (s_c - s_d)}. \tag{9}$$

Under symmetric (i.e. output-based) allocation updating, the equilibrium outcome $\{ q_c', q_d' \}$ will lie strictly above the optimal outcome on the emissions constraint line. This equilibrium outcome is associated with a level of total economic surplus that is strictly less than $S(Q^*)$ (see figure 1). Intuitively, the implicit production subsidy rebates a relatively large portion of the clean firm's explicit compliance costs. The market share of the relatively clean producer therefore increases relative to the first best benchmark. With a larger share of the market supplied by the relatively clean firm, the emissions cap can be satisfied at a higher level of aggregate output; production (and consumption) increases relative to the first best case. Appendix 2 demonstrates that the equilibrium permit price under output-based updating will be unambiguously higher than τ^*, reflecting higher supply-side marginal abatement costs.

The efficiency and distributional implications of *asymmetric* updating will depend on the ratio of the updating parameters s_c and s_d. If the implicit subsidy per unit of pollution is exactly equal across firms (i.e. implying that $\frac{s_d}{s_c} = \frac{e_d}{e_c}$), the first best level of output is achieved (see Appendix 3). However, if the implicit subsidy per unit of emissions is higher for the relatively dirty (clean) firm, the permit price must rise above (fall below) the true marginal abatement cost in order to counteract this asymmetry in compliance incentives. Note how the asymmetry in production subsidies drives a wedge between the market clearing permit price given by [9] and the true economic cost of an incremental change in supply-side emissions reductions. For example, when $s_c > s_d$, τ' does not reflect the full cost of achieving an incremental emissions reduction through reallocation of production.

In summary, contingent allocation updating distorts outcomes away from the efficient short-run equilibrium in this first best setting (except in the very special case where $\frac{s_i}{s_j} = \frac{e_i}{e_j}, \forall\ i \neq, j)$. Abatement efforts will be inefficiently allocated across producers and consumers, and across firms with different production technologies and cost structures.

2.5 Motivating the empirical exercise

General equilibrium models calibrated to specific policy contexts have been used to quantitatively estimate the potential magnitude of the distortions associated with contingent updating. In several instances, inefficiencies induced by allocation updating have been found to be economically significant (Burtraw *et al.* (2005), Jensen and Rasmussen (2000), Neuhoff *et al.* (2005)). Much of the theory literature is devoted to extending this kind of analytical exercise to more complicated, second-best settings. In cases where the implicit subsidy can be used to mitigate one or more pre-existing distortions or imperfections (such as the exercise of market power in the product market, or incomplete emissions regulation) contingent allocation updating may welfare dominate grandfathering and auctioning.

This literature is predicated on the assumption that compliance cost minimizing firms in cap-and-trade programs fully account for all permit market incentives in their supply decisions However, there are several reasons why this standard assumption might not hold in practice. First, the behavioral finance literature offers evidence to suggest that private sector decision managers may focus on information that is more readily accessible and easy to understand at the expense of information that is more opaque or that requires more resources to process (Hirshleifer, 2001; Sarin and Weber, 1993). It is presumably much easier for plant managers to translate permit prices into compliance costs per unit of production than it is to understand what a change in permit price implies for the implicity subsidy conferred by updating.

Researchers have also found evidence of gain loss asymmetry (whereby agents place more emphasis on minimizing losses versus maximizing gains) and asymmetric cost pass through in private sector decision-making (Fiegenbaum, 1990). For example, a recent study presents empirical evidence from the European carbon market that suggests firms have passed permit price increases through to customers at a different rate than price reductions (Zachmann and von Hirschhausen, 2008). To the extent that these phenomena affect firms' environmental compliance decisions, managers may discount- or ignore- the implicit production subsidy.

Third, institutional features unique to the industries subject to cap-and-trade programs may influence how firms respond to permit price incentives. For instance, in regulated wholesale electricity markets, rules governing cost recovery will play an important role in determining industry response to market-based emissions regulation. Finally, if there is any uncertainty regarding how the cap-and-trade regulation will be implemented or modified in the future, managers may discount the implicit production subsidy to reflect this regulatory risk.

In sum, whereas it is standard to assume that the explicit compliance cost and implicit production subsidy will be weighed equally in firms' production decisions, this may not be the case in practice. The obligation to hold permits to offset uncontrolled emissions may affect short-run compliance decisions differently than the implicit subsidy conferred by permit allocation updating. If firms discount- or ignore- the implicit subsidy conferred by contingent allocation updating, permit allocation incentives will not have the intended effects on market outcomes.

3 Empirical application: The NOx Budget Program

The data in this study come from a major U.S. emissions trading program: The NOx Budget Program. In 1998, the U.S. Environmental Protection Agency (EPA) determined that 23 eastern states were contributing significantly to ozone non-attainment problems. These states were issued "NOx budgets" and required to design and implement regulations that would reduce seasonal NOx emissions to budget levels. Although states had flexibility in choosing their compliance strategies, they were invited to meet their compliance obligations by joining an EPA-administered cap-and-trade program. All states accepted the invitation.

States in the NOx Budget Program (NBP) were required to accept program design features outlined in a model rule that was issued by the EPA. These features include permit trading protocols and emissions reporting standards. For instance, throughout the program, a NOx permit authorizes the holder to emit one ton of NOx during "ozone season" (i.e. May to September).[15]

[15] Compliance is only required in the spring and summer when average temperatures rise and NOx emissions

At the end of each season, all regulated source must hold sufficient permits to offset ozone season emissions.[16] There are no spatial trading restrictions in the NBP; permits are freely traded among all participating sources in all participating states.

The model rule also required standardization of intertemporal trading restrictions across participating states. Emitters cannot borrow against future allocations. Emissions in year t must be offset using permits of vintage t or earlier. Permits can be banked, although the use of banked permits is subject to a "progressive flow control" (PFC) constraint designed to discourage the excessive use of banked permits in a particular year.[17]

3.1 Permit allocation in the NBP

In the process of designing the NBP model rule, the US EPA commissioned an ex ante analysis of permit allocation design alternatives. A detailed numerical simulation model of the electricity sector was used to evaluate various allocation regimes, including grandfathering, output-based allocation updating, and fuel input-based allocation updating (US EPA, 1999)[18]. Simulation results indicated that the permit allocation design choice would appreciably affect market outcomes. Consumer electricity prices were projected to be 3.4 percent lower in an allocation updating regime as compared to grandfathering (a transfer from producers to consumers of $1.25 billion). The study also indicated that emissions leakage to neighboring, unregulated states would be reduced under allocation updating; simulated electricity production in regulated jurisdictions was 10 percent higher under updating as compared to grandfathering. With fewer emissions reductions coming from demand-side conservation/substitution and/or substitution of imports for domestic

contribute to smog formation.

[16] If a facility's emissions exceed its permit allocation, the facility must purchase additional NOx permits in the permit market. Compliance has been nearly perfect over the duration of the program; the few cases of non-compliance have been attributed to accounting errors.

[17] By law, if the number of permits in the region-wide bank prior to ozone season exceeds 10 percent of the total (i.e. program-wide) cap for that season, a non-linear discount factor is applied. The PFC ratio is computed as 10 percent of the seasonal cap divided by the size of the bank. This ratio defines the fraction of banked permits that can be used to offset a ton of permits that season. The remaining permits can be used to offset only a half ton.The discount factor is applied at the facility level. For example, if a single firm holds 100 permits and the ratio in year t is defined to be 0.5, that firm can use 50 banked permits to offset emissions in year t on a one for one basis.

[18] Because over 90 percent of emissions regulated under this program come from electricity producers, EPA analysis focused exclusively on the electricity sector. This analysis accounted for both short and long-run responses to permit allocation incentives.

production, more of the mandated emissions reductions had to come from regulated producers. Supply-side abatement costs were projected to be 18 percent higher under updating versus grandfathering, largely due to an increased market share for relatively clean- and relatively more costly- natural gas units.

Whereas many important program design features were defined at the federal level, states were ultimately given broad flexibility with regards to permit allocation design. The EPA recommended allocation updating based on heat inputs, but states were free to deviate from this recommendation.[19] Several states chose to pursue alternative approaches.

4 Data

I use data from the first four years of the NOx Budget Program (i.e. 2003-2006) and focus exclusively on electricity producers serving restructured wholesale electricity markets in the eastern United States (i.e. the New York, New England , and Mid-Atlantic or "PJM" markets).[20] Data sources and details are provided in Appendix 4.

State-level permit allocation regimes

Table 1 reports state-level NOx budgets (which were pre-determined and do not change over the study period) and information regarding state-specific permit allocation design choices. Whereas smaller states chose grandfathering (due in part to the management resources required to administer a more complex permit allocation updating process), a majority of states chose some form of contingent allocation updating based on either output or fuel inputs. Permit allocation rules vary significantly in terms of overall regime choice (i.e. grandfathering, or contingent updating), the frequency of allocations, and the basis for distributing the allowances.

[19] Fuel-based updating was chosen over output-based updating primarily because, historically, emissions regulations had been defined in terms of mass emissions per unit of heat input.

[20] Electricity generating facilities (EGUs) comprise 87 percent of the emissions sources and over 90 percent of the NOx emissions regulated under the NBP (EPA, 2007). EGUs regulated under the NBP operate in a variety if electricity markets. Whereas units in the sample supply restructured wholesale electricity markets, other facilities in the program are rate-regulated producers serving vertically integrated, economically regulated electricity markets, while other units are owned and operated by public entities and operate on a non-profit basis. Production at rate-regulated plants and public entities are more centrally coordinated and influenced by an array of economic, regulatory, and institutional factors. I choose to focus on restructured electricity markets because EGUs in these markets are more likely to have short-run objectives consistent with profit maximization.

Unit-level operations and attributes

Table 2 summarizes some important unit-level operating characteristics by permit allocation regime. Two operating attributes that will be particularly relevant to this analysis are the NOx emissions rate (i.e. pounds of NOx emitted per MWh electricity produced) and heat rate (i.e. btus of fuel burned per kWh of electricity produced).[21] It is fairly standard in the empirical literature to treat these unit-specific performance parameters as immutable features of the production technology. However, emissions rates and heat rates can be affected by operating decisions made by the plant manager, including the choice of fuel characteristics, utilization rates, and combustion tuning. Purely exogenous factors (such as ambient temperature) can also play a role.

To construct unit-specific summary measures of these operating characteristics, separate seasonal regression equations are estimated for each unit. This estimation exercise, described in more detail in Appendix 5, obtains unit-specific, season-specific point estimates of emission rates and heat rates under average operating conditions.[22] Capacity-weighted summaries of these estimates are presented in Table 2. The support of the distributions overlap considerably across allocation regimes, which facilitates an empirical comparison of short-run production decisions made by similar units facing different permit allocation incentives.

The table also reports summary statistics for plant operating capacities and ramp-up rates (i.e. the rate at which a unit can increase production in one hour, expressed as a function of total capacity). These characteristics, discussed in more detail in Appendix 4, will also be relevant in the proceeding analysis of firms' response to permit allocation incentives. .

Emissions permit prices

NOx permits are actively traded in a liquid permit market.[23] Table 3 reports average spot NOx permit prices by vintage (in nominal dollars per ton). NOx permit prices fell over the study period, largely due to abatement costs that proved to be lower than anticipated, and lower than

[21] A unit's heat rate measures the efficiency with which the unit transforms fuel into electricity. The lower the heat rate, the more fuel efficient the generator.

[22] Specifications that allow rates to vary across years are also estimated.

[23] In 2007, the volume of "economically significant" immediate settlement trades (i.e. trades between versus within firms) reached 247,000 tons (EPA 2008).

expected temperatures in the early years of the program.[24] This table also helps to illustrate the effect of the progressive flow control (PFC) constraint on permit prices. As early as 2003, permit market participants correctly anticipated that the PFC constraint would start to bind in 2005. This explains the large vintage 2004/2005 spread in 2003 and 2004.[25]

Compliance costs and production incentives

To estimate the cost of purchasing permits to offset the emissions associated with generating a MWh of electricity, each unit's NOx emissions rate is multiplied by the NOx permit price. Table 4 summarizes these unit-level cost estimates. On average, explicit compliance costs amount to a 7 percent increase in total variable (i.e. fuel, operating and maintenance) costs.[26] However, among units with particularly high emissions rates, this increase can exceed 40 percent.

Estimating the implicit production subsidies conferred by contingent updating is more complicated. The size of the production subsidy varies with state permit allocation rules, state-specific NOx budgets, annual production levels, and unit-level heat rates (in input-based updating regimes). Individual states allocate their respective NOx "budgets" (listed in Table 1) using formulas of varying complexity.

For each unit, an estimate of the number of future permits earned per unit of current production is constructed using the corresponding state budget E_s, average ozone season production (or fuel input in states that have adopted heat input based updating) aggregated across NBP sources in the state, and the specific details of states' permit allocation updating protocols. For example, if NOx permits in state s are allocated based on the average heat input in the preceding L years, the effect of an incremental increase in current production at firm i in year t on future permit entitlements is assumed to be $h_i\left(\frac{E_s}{H_{st}}\right)$,where h_i measures the fuel inputs required to generate a unit of output at unit i, and H_{st} measures the total quantity of fuel inputs used by NBP sources

[24] Evolution Markets LLC provides informative monthly analyses of the NOx Budget Program permit market.

[25] In years when the PFC constraint binds, banked permits trade at a considerable discount. The PFC ratio was 0.25 and 0.27 in 2005 and 2006, respectively. In both years, permits were used to offset emissions at a discounted rate (4,168 and 1,950 permits in 2005 and 2006, respectively). In March of 2005, the EPA released its new Clean Air Interstate Rule (CAIR), intended to subsume the NBP in 2009. CAIR eliminated progressive flow control.

[26] Unit-specific estimates of variable fuel operating costs are obtained by multiplying the unit-level heat rate (see above) by the corresponding fuel price. Estimates of unit-level variable, non-fuel operating and maintenance costs (not including environmental compliance costs) are obtained from Platts.

in state s over the course of the ozone season in year t. An important-and plausible- assumption is that firms take the size of the per-MWh subsidy as given.[27] Column 3 of Table 4 summarizes these estimated subsidies, in terms of future permits allocated, per MWh of electricity generated.

In present value dollar terms, the estimated implicit subsidy conferred under this input-based updating regime is:

$$s_{it} = \sum_{l=1}^{L} \frac{\delta_i(l)}{L} \tau_l \left(\frac{h_{it}}{H_t} E_s \right), \tag{10}$$

where τ_l is the expected permit price in l years and $\delta(l)$ is the discount rate applied to benefits accruing l years in the future.[28] In the final column of Table 4, net compliance costs (i.e. explicit compliance costs less the implicit subsidy per MWh) are summarized.[29] These incentives vary significantly across facilities. Notably, for several units in allocation updating regimes with relatively low emissions rates, the estimated implicit subsidy exceeds the estimated explicit compliance cost such that the net effect of the NBP on variable operating costs is negative.

5 Empirical framework

The unique design of the NBP provides several potentially useful sources of variation. First, the delegation of permit design to state-level agencies has yielded significant interstate variation in permit allocation rules and related incentives. From a research design standpoint, permit allocation design features would ideally have been randomized across electricity producers. Although states' choice of permit allocation regime was not random, interstate variation in permit allocation design is arguably exogenous, in an econometric sense, to firms' short-run production decisions. State level permit allocation design decisions were determined by a variety of factors, including the institutional capacities of the implementing agency and the preferences of politically powerful

[27] A firm with a dominant market position would want to account for the fact that increasing its fuel consumption would increase H_{st} and thus decrease the size of the subsidy it received for all of its production. Thus we would expect the perceived subsidy would be decreasing in market share.

[28] To construct an estimate of these implicit subsidies, the futures price of permits issued l years in the future is used to estimate $\delta_i(l)\tau_l$. In cases where permits did not trade far enough into the future, the market price for the permit vintage farthest in the future was applied to all subsequent vintages..

[29] Rather than assume an arbitrary discount rate, these calculations present undiscounted estimates of the implicit production subsidy. Plant managers presumably discount the value of future permit allocations, so these estimates should be interpreted accordingly.

constituents. Conditional on pre-determined industry and production technology characteristics, the factors that shaped a state's choice of allocation regime should not directly impact unit-level short-run production decisions.

Second, the seasonal nature of the program's compliance requirements generates useful intertemporal variation. There is considerable overlap in the distribution of hourly load levels, and other observable market conditions across ozone season and off-season (see Table 5). This makes it possible to observe unit-level production decisions in hours that differ in terms of ozone compliance requirements, but are otherwise similar.

Third, a subset of NOx emitting producers supplying the New England electricity market are exempt from the NBP for meteorological reasons. Most counties in Maine, Vermont, and New Hampshire were already in compliance with the ozone standard. Prevailing wind and weather patterns ensure that emissions from point sources in these states do not contribute significantly to U.S. non-attainment problems. The distributions of operating characteristics that determine short-run production decisions in the exempt and NBP regulated sub-populations overlap considerably (see Table 2), making these exempt units a potentially useful control group.

Finally, a majority of states chose to adopt the EPA's recommended permit allocation approach: heat input based updating. Under this regime, the production subsidy varies significantly with fuel efficiency. Input-based updating thus generates inter-facility, intra-market variation in production incentives that is independent of variation in explicit compliance costs per unit of production.[30]

5.1 Modeling the data generating process

The model developed in section 2 serves as a good starting point for a model of the data generating process. However, this model needs to be modified to more accurately reflect the process that generated these data. To begin, unit-level capacity constraints (which routinely bind in these electricity markets) are imposed and unit-level marginal operating costs are assumed to be approximately constant. Having made these two modifications, the model closely resembles those

[30] The correlation coefficient between unit-level fuel efficiency measures and NOx emissions rates.is 0.69.

used to simulate wholesale electricity market outcomes in competitive benchmark analysis (see, for example, Borenstein, Bushnell, and Wolak, 2002 and Wolfram, 1999) and environmental policy simulations (examples include US EPA, 1999 and Burtraw, Palmer and Kahn, 2005).

This model can be used to make predictions about short-run, unit-level electricity supply decisions. Figure 2 helps to illustrate these predictions. The vertical axis measures unit-level capacity factor. The horizontal axis measures wholesale electricity price. The model predicts that profit maximizing producers will follow an on-off strategy, producing at full capacity whenever price exceeds a reservation price set equal to the unit's constant variable operating cost. In the off-season (i.e. when firms are not required to hold permits to offset emissions) this reservation price is equal to the fuel costs, labor costs, and other variable costs incurred per unit of electricity generated. During ozone season, this reservation price increases by an amount equal to the unit-specific net compliance cost per MWh. Because the average net effect of the NBP on variable operating costs is substantially higher among units operating in grandfathering regimes (see table 4), we should expect that the introduction of the NBP will have a more significant average impact on the reservation prices of producers in grandfathering regimes.

Are observed production decisions consistent with these predictions? Figure 3 is generated using a small subset of the data collected from a representative unit over a short (three day) period in the ozone off-season.[31] The left panel plots capacity utilization and hourly wholesale electricity prices over these 96 hours. The horizontal line represents the estimated marginal off-season operating cost of $49/MWh specific to this unit and time period (i.e. the reservation price). The right panel plots capacity factor as a function of the wholesale electricity price less variable operating costs. The thick black step function plots the relationship predicted by the model. The thin S-shaped curve is a local polynomial smooth of the observed data. Note that observed hourly production decisions at this unit deviate systematically from the predictions of the simple, static model.[32]

[31] I chose a period in the ozone off-season in which the wholesale electricity price was vascillating around this unit's theoretical reservation price (i.e. the prevailing fuel price multiplied by the unit's fuel efficiency rating plus variable, non-fuel operating costs).

[32] These supply decisions are observed in the ozone off-season. The average net compliance cost incurred by this unit during ozone season is estimated to be $3.16. The thick red line illustrates how the introduction of the emissions trading program should, in theory, affect the unit's hourly production decisions.

Figure 4 conducts a similar exercise using the complete data set. The vertical axis measures capacity factor. The horizontal axis measures price less variable operating costs (not including NBP compliance costs). The solid line in each panel plots the local mean smooth of hourly, unit-level capacity utilization rates on hourly, unit-specific price-cost margins in the ozone off-season.[33] These functions are generated separately for grandfathering and contingent updating regimes, respectively. The broken lines plot the same relationships using data from ozone season (i.e. May to September) when units are required to hold permits to offset their NOx emissions.

Taken together, these figures are only partly consistent with theory. The short run supply functions generated using data from ozone season lie to the right of their off-season counterparts as expected. However, based on these figures alone, it is impossible to tell if this NBP-induced shift in the supply curve is more significant in the right panel . Finally, observed production decisions deviate systematically from the predicted on-off production protocol. These units are observed operating at less than full capacity in a majority of hours. Plant managers appear willing to operate in hours when prices fall below marginal operating costs, and appear slow to respond when the prices rise above cost.

Much of this behavior can be attributed to technological and system operating constraints that are omitted from the model. At the unit level, ramping limits, start up costs, minimum run times, and other intertemporal operating constraints can significantly affect how a plant responds to changing market conditions. At the system-level, transmission constraints, system-security requirements, and other operating protocols can affect which units get called upon to run in a given hour.

In fact, the true data generating process involves a complex, multi-period optimization problem. This so-called "unit commitment" problem (i.e. the dynamic scheduling of electricity production over hours in a day.) has been extensively analyzed in the operations research and power systems literature. The problem is difficult to solve because of its large dimension, non-linearity, and large number of constraints (Sheble and Fahd 1990). One formulation of the unit commitment

[33] To generate these figures, I use an Epanechnikov weight function and a rule-of-thumb bandwidth estimator. The smooth is evaluated at 50 points. Price cost margins are calculated by subtracting a unit's fuel costs and non-fuel variable operating costs per MWh from the hourly real-time wholesale electricity price.

problem, introduced in Appendix 5, helps to convey the complexity of the dynamic optimization faced by plant managers and system operators each day.

Previous work has demonstrated the perils of ignoring unit commitment idiosyncrasies when simulating short-run wholesale electricity market outcomes in the context of competitive benchmark analyses (Harvey and Hogan, 2002; Mansur, 2008). Ignoring unit commitment constraints will also be problematic in this setting. Observed departures from the static model of profit maximization are likely correlated with factors that determine a plant's response to permit allocation incentives.

Estimating a fully structural econometric model of unit commitment would be computationally intensive and would require additional assumptions about the nature of both the system-level and unit-level operating constraints and protocols. Given the limitations of the available data, these assumptions would be ad hoc and untestable.

In what follows, I adopt a slightly less ambitious strategy. I focus on one dimension of the larger unit commitment problem: the decision to begin operating conditional on being inactive. An exclusive focus on this participation margin facilitates the derivation a reduced form that can be implemented empirically as a discrete choice problem. Appendix 7 summarizes a complementary, more descriptive approach to analyzing how the statistical relationship between unit-level supply decisions, NOx permit prices varies with observable unit-level characteristics (such as emissions rates and implicit production subsidies) when other observable determinants of the supply decision are flexibly controlled for.

5.2 A reduced form model of the participation decision

Consider the unit commitment problem faced by a single electricity generating unit (see Appendix 5). At the beginning of period t, the manager of unit i must choose an output level to maximize profit π_{it}.[34] Let q_{it} measure the output level at unit i at the beginning of hour t. The control

[34] In the restructured electricity markets considered in this paper, market participants submit bids to an independent system operator (ISO). Unit-level production activities are coordinated via a two-settlement market system. A day ahead forward market schedules resources and determines hourly prices for the following day; a balancing market ensures that supply meets fluctuating demand in real time. Once generators have submitted their supply bids, independent system operators identify unit commitment schedules to minimize the cost of meeting electricity demand subject to thousands of unit-level and system-level operating constraints. I assume that hourly supply

variable d_{it} measures the change in output at unit i in hour t. The transition equation that determines the evolution of the state variable q_{it} over time is thus $q_{it} + d_{it} = q_{it+1}$. The choice of d_{it} is constrained by a suite of operating constraints. Let D define the decision space. The set of possible production level changes available to unit i in hour t, D_{it}, will depend on time invariant parameters of the operating constraints Γ_i and the state variable q_{it}.

Profits earned from the sales of electricity generated at unit i in hour t are:

$$\begin{aligned} \pi_{it} &= (P_{it} - c_i - \theta^t \tau_t e_i + \theta^s \ \tau_{t+s} s_i)(q_{it} + d_{it}) - U_i y_{it} - F_i, \ d_{it} \in \{D_{it}(q_{it}, \Gamma_i)\} \qquad (11) \\ &\equiv F(q_{it}, d_{it}, X_{it}), \end{aligned}$$

The NOx permit price is τ. Unit-level emissions rates and implicit subsidies (in terms of future permits earned per unit of electricity generated) are represented by e_i and s_i, respectively. The θ parameters are included to allow firms to weigh explicit compliance costs and the implicit subsidy asymmetrically in their production decisions. The binary variable y_{it} equals 1 if the unit turns on in hour t; otherwise $y_{it} = 0$. Start-up costs and fixed costs are represented by U_i and F_i, respectively. The X_{it} matrix includes state variables observed by both the plant manager and the econometrician, including fuel prices, permit prices, and electricity prices.

Because most electricity generating units are incapable of responding quickly to changing market conditions, production levels in one hour will constrain production possibilities in future hours. Let H_i denote the relevant time horizon (measured in hours) for the unit commitment problem solved by unit i. Let j index the participation choice situation and t index the hours relevant to the participation decision. The choice of $d_{i0} \in \{D_{i0}(q_{i0}, \Gamma_i)\}$ has repercussions for $\{D_{it}(q_{it}, \Gamma_i)\}$ for $t = 0...H_i$. [35]

The plant manager's objective is to maximize multi-period profits $\Pi(q_{i0}, X_{ij}, d_i) = \sum_{t=0}^{H_i} F(q_{it}, X_{ij}, d_{it})$ subject to the transition function $T(q_{it}, d_{it})$ and the constraint sets $D_{ij}(q_{ijt}, \Gamma_i)$. Within this framework, theory makes clear predictions about when an inactive unit should start producing. A profit maximizing manager will choose to incur the costs of initiating operations if the revenues less

decisions are controlled by plant managers insofar as they can submit bids to the independent system operator to achieve their desired production levels.

[35] Among the most nimble units, $H = 0$ and the production decision reduces to the static benchmark model.

costs from doing so exceed the revenues less costs associated with remaining out of the market. Conditioning on $q_{it} = 0$ simplifies the derivation of an estimable reduced form model.

I define choice specific value functions $v(y_i, X_{ij})$ to capture the expected profits associated with participation choice:

$$
\begin{aligned}
v(1, X_{ij}) &= E_0[F(1, d^*_{i0}(1), X_{ij}) + \sum_{t=1}^{H_i} F(q_{it}, d^*_{it}(1), X_{ij}) + \eta^1_{ij}] \\
v(0, X_{ij}) &= E_0[F(0, 0, X_{ij},) + \sum_{t=1}^{H_i} F(q_{it}, d^*_{it}(0), X_{ij}) + \eta^0_{ij}]
\end{aligned} \tag{12}
$$

The X_{ij} matrix includes all state variables relevant to the participation choice made by unit i in choice situation j. These include fuel prices and permit prices (which are constant over the time horizon H_i), the real time electricity price in hour 0 and the day ahead forecast electricity prices in hours $1..H_i$. The optimal production choice in hour t conditional on the initial participation choice y is $d^*_t(y)$. A decision specific shock η^y captures the effects of unobserved factors affecting expected returns to either starting to produce or remaining inactive. These factors could include plant efficiency shocks, unscheduled outages, or optimization errors. I assume these shocks are additive and independently distributed N(0,σ).

Let the latent variable y^*_{ij} measure the difference in these conditional value functions. To complete the motivation of the econometric model, I derive the following reduced form of the decision rule that is linear in electricity prices, permit prices, and marginal operating costs:

$$
y^*_{ij} = E_0[F(0, d^*_0(1), X_j) + \sum_{t=1}^{H} F(q_t, d^*_t(1), X_j) + \upsilon^1_j] - E_0[F(0, 0, X_j,) + \sum_{t=1}^{H} F(q_t, d^*_t(0), X_j) + \upsilon^0_j] \tag{13}
$$

$$
= \alpha_i + \sum_{t=0}^{H} \beta^P_{ijt} P_{it} - \beta^c_{ij} c_{ij} - \beta^e_{ij} \tau_j e_i + \ \beta^s_{ij} \tau_j s_i + \epsilon_{ij}. \tag{14}
$$

The fixed effect α_i captures unit-specific start-up costs, fixed operating costs, and other time-invariant factors influencing the participation decision. The set of electricity price coefficients β^P_{ijt} , $t = 0...H_i$, capture the hour by hour differences in optimal output levels (conditional on the participation decision made in period 0) for a particular unit and choice situation: $\beta^P_t \equiv$

$(q_t(1)+d_t^*(1))-(q_t(0)+d_t^*(0))$. I define the parameter Δ_{ij} to capture the effect of the participation decision on total production over the time horizon H_i in choice situation j: $\Delta_{ij} = \sum_{t=1}^{H}(q_{ijh}(1) + d^*_{ijh}(1) - q_{ijh}(0) - d^*_{ijh}(0))$. These Δ_{ij} parameters vary both across units and within units across choice situations characterized by different electricity market conditions, fuel prices, etc. Given the assumed structure of the underlying economic model, the reduced form parameters β^c, β^e, and β^s coefficients are equal to Δ_{ij}, $\theta_i^t\Delta_{ij}$, and $\theta_i^s\Delta_{ij}$, respectively. The random state variable ϵ_{ij} is assumed to be normally distributed (arising from the difference between η_{ij}^0 and η_{ij}^1).

In [14], the contemporaneous permit price τ_t is used to proxy for firms' expectations regarding future permit prices in [11]. This simplifying assumption will solve a multicollinearity problem that arises from the fact that spot permit prices and futures prices are highly correlated. However, firms presumably use a non-zero discount rate when valuing benefits accruing in future periods, and table 2 illustrates how futures prices can deviate from spot prices in this permit market. Consequently, the current permit price is likely to be an over-estimate of expected future permit prices. This will have implications for how the estimation results are interpreted.

The observed binary choice variable y_{ij} serves as an indicator that the latent value $y_{ij}^* > 0$: $y_{ij} = 1\{y_{ij}^* \geq 1\}$.The probability that an inactive unit i facing choice situation j will begin to operate is given by:

$$\Pr(y_{ij} = 1|X_{ij}, \Gamma_i) = \Phi\left(\alpha_i + \sum_{t=0}^{H} \beta_{ijt}^P P_t - \beta_{ij}^c c_{ij} - \beta_{ij}^e \tau_j e_i + \ \beta_{ij}^s \tau_j s_i\right), \tag{15}$$

where Φ is the cumulative distribution function of the standard normal distribution.

One disadvantage of this reduced form estimation approach is that all unit-hours in which a unit is initially operating are dropped from the data set used to estimate the model. This amounts to omitting more than 80 percent of observations at coal-fired units, approximately 25 percent of observations at natural gas fired units, and 14 percent of hourly observations at oil-fired units. Appendix 7 describes an alternative estimation approach that uses the complete data set to assess the extent to which observed statistical patterns are qualitatively consistent with theoretical predictions.

6 Estimation

The parameter values in equation [14] are likely to vary significantly across units with different production technologies, constraints, and system operating protocols. It will be important to capture as much of this cross-sectional heterogeneity as possible. Each individual unit is observed making production decisions over several thousand hours. Electricity prices, fuel prices, and permit prices vary significantly across time. It is therefore possible, in principle, to consistently estimate average values of β^P_{ijt} and β^c_{ij} separately for each unit. However, variation in unit-specific explicit compliance costs $\tau_j e_i$ that is independent of variation in unit-specific implicit subsidies $\tau_j s_i$ exists only in very limited quantities[36] An empirical investigation of permit allocation incentives must therefore exploit cross-sectional variation in operating characteristics and state-level permit allocation design choices.

Estimation could proceed in one step using a fully pooled probit model, but this is cumbersome because it involves thousands of interactions between variables in X_{it} and unit-specific dummy variables. Estimating the model in two steps makes efficient use of these hierarchical data without incurring large computational costs. The main disadvantage is that contemporaneous correlation across the unit-level equations cannot be accommodated.

6.1 *First stage of estimation*

In the first stage, unit-specific probit equations are estimated using hourly, unit-level data:

$$y_{ij} = 1\{\alpha_i + \sum_{t=0}^{H} \beta^P_{it} P_t - \beta^c_i c_{ij} - \beta^1_i \tau_j \cdot D\_OZ_j + \epsilon_{ij}\}.$$

The binary variable D_OZ_j equals one during ozone season and zero otherwise. The permit price τ_j and operating costs c_{ij} are those observed over the H hours associated with choice situation j. Real-time and day ahead hourly locational marginal prices were matched to each unit using

[36] There is some within-unit variation in emissions rates and heat rates (and thus per-MWh compliance costs) across years due to retofits and combustion modifications that occur during the study period. This variation is extremely limited as compared to the significant cross-sectional variation in emissions rates and implicit production subsidies.

detailed hourly data made available by the Independent System Operators. Day-ahead prices were used to proxy for expected future prices. These location-specific marginal electricity prices may be endogenous if unobserved supply shocks affect both market clearing prices and unit-level supply decisions. To address potential endogeneity concerns, ISO-specific hourly demand forecasts are used to instrument for these location specific electricity prices.Day ahead demand forecasts should be independent of unobserved supply shocks but highly correlated with realized demand conditions and thus electricity prices in a given hour.

6.2 *Second stage of estimation*

In the second stage, the following equation is estimated using a feasible generalized least squares (FGLS) estimator:

$$\widehat{\beta_i^1} = \beta^0 + \beta^e e_i + \beta^s s_i + \upsilon_i. \tag{16}$$

Emissions rates at units exempt from the NOx Budget Program are set to zero. This e_i variable is demeaned such that the constant term β^0 captures the average relationship between an incremental change in the NOx permit price and the average latent value y_i among units with average NOx emissions rates.[37] The β^e coefficient captures the interaction between variation in permit prices and unit-level emissions rates. The β^s coefficient captures the interaction between variation in permit prices and the estimated subsidy for those units in contingent updating regimes. In more flexible specifications, all coefficients are allowed to vary with observable unit characteristics.

The second stage residuals contain two components: a sampling error component (i.e. the difference between the true value of the estimated dependent variable and the true value) and the idiosyncratic variation that would have obtained regardless of whether the dependent variable was estimated or observed directly. If the sampling variance differs across units, this component will be heteroskedastic. To address this issue, a FGLS estimator is used to incorporate information about

[37] I treat unit-level emissions rates as fixed. No attempt is made to account for the stochastic properties of these unit-specific operating parameters (see Appendix 4). Future work will explore alternative approaches to incorporating this variation. For example, Joskow and Schmalensee (1985) demonstrate a consistent (adjusted least-squares) technique for using estimated plant operating characteristics as independent variables in crosssection regression analysis.

the variance structure resulted in the first stage estimation (Hanushek, 1974).[38] The weighting matrix used in the second stage is given by:

$$\widehat{\Omega} = \widehat{G} + \widehat{\sigma^2} I, \tag{17}$$

$$\widehat{\sigma^2} = \frac{\sum_i \varepsilon_i^2 - \sum_i \upsilon_i^2 + tr(Z'Z)^{-1}\widehat{G}Z}{N-k} \tag{18}$$

where $\widehat{G}$ is the variance covariance matrix from the first stage, ε_i are the first stage standard errors of the estimated coefficient that serves as the dependent variable in the second stage, υ_i are the residuals from an OLS estimation of the second stage, k is the number of regressors in the second stage and Z is the matrix of regressors. This estimator weights more precise first stage estimates more heavily, but only to the degree that sampling error is an important component of the overall second stage residual. Standard errors are also clustered at the facility level to account for the fact that the idiosyncratic component of the error term may be correlated among boilers located at the same facility.

Interpretation of these second stage estimates is complicated by two identification issues. First, parameters are identified only up to a scale factor (Maddala 1983). Consequently, comparisons of coefficient estimates across units confound the magnitude of the regression coefficients with residual variation. The second issue pertains to the structure of the reduced form estimating equation. Permit price coefficients are also confounded with the Δ_{ij} parameters.[39] If responsiveness to permit allocation incentives varies across units with different emissions rates and/or different implicit subsidies in ways that are correlated with the unobserved residual variances from the first stage and/or the Δ_{ij} parameters, coefficient estimates may be biased. In order to identify the relative effects of the explicit compliance costs and implicit production subsidies conferred by

[38] A more common approach to adjusting standard error estimates when the dependent variable is estimated involves weighting second stage observations using the inverse of the estimated standard error of the dependent variable (Saxonhouse, 1976). However, this assumes that the total residual is heteroskedastic (versus just the component that is explained by sampling error).If the variance due to sampling error accounts for a relatively small fraction of the total residual variance in the second stage, this reweighting can generate misleading standard error estimates.

[39] The ratio of the marginal cost and NOx permit price coefficient is identified in principle; scale parameters and Δ parameters cancel out. One approach would involve using this ratio as the dependent variable in the second stage in order to overcome these identification problems. However, this is difficult to implement in practice. Because the distribution of both coefficients include zero, some estimates of this ratio can be infinitely large.

emissions trading programs, I will need to assume that these scale parameters and Δ_{ij} parameters are uncorrelated with the factors that determine firms' response to permit allocation incentives. In this section, I present some alternative specifications and indirect tests which are generally supportive of this assumption.

6.3 *Results*

Table 6 summarizes estimation results from the first stage. The preferred specification assumes a time horizon of six hours.[40] This table also reports estimation results from specifications that assume shorter and longer time horizons. Hourly demand forecasts are used as instruments for real time and day ahead electricity prices. The preferred specification also includes year and day of week dummies (not reported). As expected, estimated marginal cost coefficients are negative on average. The average NOx price coefficient estimate is also negative and highly variable across units. Estimated wholesale electricity price coefficients are of the same order of magnitude. Because contemporaneous spot and day ahead electricity prices are highly correlated, the individual price coefficients are difficult to interpret.

Estimation results from the second stage are presented in Table 7. The dependent variable is the estimated β_i^1 coefficients from the first stage. The most restrictive specification includes only a constant and demeaned NOx emissions rate. The constant term captures, among other factors, the effect of a change in the NOx permit price on the latent value y^* among units with average NOx emissions rates. Somewhat surprisingly, I fail to reject the null hypothesis of zero average effect. We should expect y^* - and thus participation decisions- to be negatively affected by a change in the permit price. The NOx emissions rate coefficient is negative and statistically significant, as expected. This implies that the effect of a change in the permit price on the latent y^* value is significantly more (less) negative among units with relatively high (low) emissions rates (and therefore relatively high (low) compliance costs.

A less restrictive specification (2) allows the NOx price coefficient to vary with the implicit

[40] The choice of a six hour time horizon was informed by an informal analysis of ramping constraints. Unit-specific ramp rates were crudely estimated by looking at hour-to-hour changes in production, expressing these changes as a percentage of installed capacity, and defining the ramping constraint to be the 90th percentile hourly change. Because the vast majority of estimated ramp rates exceed 16 percent, setting H.

subsidy introduced by contingent allocation updating regimes. The emissions rate coefficient increases slightly in absolute value and remains highly statistically significant. The coefficient on the subsidy variable is positive as expected, but imprecisely estimated.

Several alternative specifications were estimated so as to allow the NOx rate and subsidy coefficients to vary with observable unit-level characteristics. One of the specifications that best fit the data allows these coefficients to vary with installed capacity (column 3). Results indicate that production decisions of larger emitting producers are more responsive to a change in emissions permit prices (and thus compliance costs) as compared to smaller producers. With the addition of these interaction terms, the subsidy effect can be more precisely estimated. This suggests that larger producers are also more likely to respond to the implicit subsidy.

Columns (4), (5), and (6) explore the robustness of these results to varying assumptions about the endogeneity of first stage covariates and the assumed time horizon. Column (4) reports results obtained when electricity prices are assumed to be exogenous in the first stage. Columns (5) and (6) report results under varying assumptions about the time horizon. The NOx rate coefficient is negative and highly statistically significant across all specifications. The statistical significance of the implicit subsidy and the interaction terms are somewhat less robust to these specification changes.

In order to use these results as a basis for causal inference regarding the impacts of unit-level emissions rates and subsidies on the relationship between permit prices and short-run participation decisions, I must assume that the residual variance and the Δ_{ij} parameters are distributed independently of permit allocation incentives. If this is not the case, strong correlations between NOx price coefficients estimated in the first stage and emissions rates or implicit subsidies could be spurious.[41] To investigate this possibility, additional specifications are estimated. Although the Δ_{ij} parameters and residual variances are not observable, variables that are plausibly strongly correlated with these factors are observed in the data. For example, fast ramp rates should be associated with smaller Δ_{ij} parameters. Larger operating capacities should be associated with

[41] For example, if the Δ_{ij} parameters are systematically smaller among units with higher emissions rates, this would result in a negative NOx emissions rate coefficient in the second stage that has nothing to do with a causal relationship between high emissions rates and the nature of producers' response to changing NOx permit prices.

larger Δ_{ij} parameters, all else equal.

Columns (7) and (8) constitute indirect tests of these independence assumptions. Ramp rates and unit capacity, two variables plausibly correlated with the Δ_{ij} parameters, are added to the model in (7). Although both have the expected sign, only the ramp rate coefficient is statistically significant. The other coefficient estimates are not substantially affected.

The error term captures, among other things, unobserved differences in system operations. Residual variances could be expected to vary across regional electricity markets with different dispatch procedures and protocols. Indicator variables for the New York and New England market are added to the model in (8). The New York coefficient is positively and weakly significant. This could be a result of smaller residual variance among units supplying New York, or it could be capturing the average effect of the state's input-based allocation updating regime. The other coefficient estimates are not significantly affected. Taken together, these results provide weak support for the aforementioned independence assumptions.

In sum, these results suggest that electricity suppliers of average size do respond to both explicit compliance costs and implicit production subsidies associated with the introduction of this cap-and-trade program. The NOx emissions rate coefficient is negative and strongly statistically significant across all specifications. The point estimate of the effect of the implicit subsidy is consistently positive and smaller in absolute value. Taken at face value, these point estimates are roughly consistent with a discount rate of 10-15 percent (permit allocations are updated with a three to four year lag on average). Based on these admittedly noisy estimates, the hypothesis that plant managers ascribe equal weight to the implicit subsidy and the explicit compliance costs in their short-run production decisions cannot be rejected.

7 Conclusions

Policymakers, industry representatives, and other stakeholders are increasingly interested in understanding how the design of permit allocation protocols can affect permit and product market outcomes. A growing theoretical literature offers insights into how firms should respond to different permit allocation rules, including "updating" regimes that simultaneously penalize emissions while rewarding production. Empirical evidence has thus far proved elusive. This paper analyzes the production decisions of hundreds of electricity generators in a large emissions trading program that is particularly well suited to a study of the short-run impacts of permit allocation design.

A simple partial equilibrium model demonstrates the theoretical, short-run implications of contingent updating vis a vis more traditional permit allocation designs (i.e. auctioning and grandfathering) and serves to motivate the empirical work. Overall, no strong statistical relationships between permit market price variation and the overall average short-run production patterns. Among larger producers, statistical patterns are generally consistent with the standard theory regarding how permit allocation incentives should affect market outcomes.

A more rigorous empirical strategy imposes more structure so as to provide a basis for causal inference and hypothesis testing. For the sake of tractability, I focus exclusively on the participation margin of the unit commitment process that presumably generates the observed short-run supply decisions. I derive and estimate a reduced form model of electricity producers' decisions to resume operations conditional on being inactive. Among sellers of average production capacity, supply decisions appear to respond to both the explicit emissions tax and the implicit production subsidy conferred by this emissions trading program. I fail reject the hypothesis that negative emissions incentives and positive production incentives conferred by allocation updating are weighed equally in short-run production decisions. When I do not condition on production capacity, I cannot statistically distinguish the overall average response of short-run production decisions to variation in the permit price from zero.

This analysis is not without its limitations. Precise estimation of the market-level impacts and associated welfare implications of the permit allocation design decisions I analyze is beyond the scope of this paper. A burgeoning literature endeavors to predict welfare impacts of permit allocation design choices in a range of counterfactual policy contexts using detailed, theory-based simulation models. Although the results of these studies cannot be confirmed or refuted based on evidence presented here, the empirical findings lend support to some pivotal assumptions underpinning these simulation exercises.

Finally, an important motivation for this study is to inform future permit market design decisions, particularly those that will determine how hundreds of billions of dollars worth of carbon credits will be allocated under proposed Federal climate change legislation. The paper provides empirical evidence that firms are responding to the relatively modest permit allocation incentives conferred by the NOx Budget Program. We might expect this response to carry over to a larger greenhouse gas emissions trading program where the stakes will likely be significantly higher. However, the regional emissions trading program analyzed in this paper and proposed Federal program designed to limit greenhouse gas emissions differ along a number of important design dimensions. Additional research is needed before we can definitively anticipate how firms will respond to permit allocation incentives in the next generation of cap-and-trade programs.

References

Bernard, A. L., Fischer, C. & Fox, A. K. (2007), 'Is there a rationale for output-based rebating of environmental levies?', *Resource and Energy Economics* **29**(2), 83–101.

Bohringer, C. & Lange, A. (2005), 'Economic implications of alternative allocation schemes for emission allowances', *The Scandinavian Journal of Economics* **107**(3), 563–581.

Borenstein, S., Bushnell, J. B. & Wolak, F. A. (2002), 'Measuring market inefficiencies in california's restructured wholesale electricity market', *American Economic Review* **92**(5), 1376–1405.

Bunn, D. & Fezzi, C. (2007), 'Interaction of European Carbon Trading and Energy Prices', *SSRN eLibrary* .

Burtraw, D., Palmer, K., Bharvirkar, R. & Paul, A. (2002), 'The effect on asset values of the allocation of carbon dioxide emission allowances', *Electricity Journal* pp. 51–62.

Burtraw, D., Palmer, K. & Kahn, D. (2005*a*), Allocation of CO2 emissions allowances in the regional greenhouse gas cap-and-trade program, Discussion Papers dp-05-25, Resources For the Future.

Burtraw, D., Palmer, K. & Kahn, D. (2005*b*), CO2 allowance allocation in the regional greenhouse gas initiative and the effect on electricity investors, Discussion Papers dp-05-55, Resources

For the Future.

Bushnell, J. & Chen, Y. (2009), Regulation, allocation, and leakage in cap-and-trade markets for co2, Technical report, EI@Haas WP-183, Washington, DC.

Commission, C. P. U. & the California Energy Commission (2008), Joint california public utilities commission and california energy commission staff paper on options forallocation of ghg allowances in the electricity sector, Technical report, R.06-04-009 and D.07-OIIP-01.

Cramton, P. & Kerr, S. (2002), 'Tradeable carbon permit auctions: How and why to auction not grandfather', *Energy Policy* **30**(4), 333 – 345.

Dinan, T. & Rogers, D. (2002), 'Distributional effects of carbon allowance trading: How government decisions determine winners and losers', *National Tax Journal* **2**, 199–221.

Duxbury, D. & Summers, B. (2004), 'Financial risk perception: Are individuals variance averse or loss averse?', *Economics Letters* **84**(1), 21 – 28.

Environmental Protection Agency, U. (1999), Economic analysis of alternate methods of allocating nox emission allowances, Technical report, ICF Consulting.

Fell, H. (2009), 'Eu-ets and nordic electricity: A cvar analysis', *Energy Journal* .

Fiegenbaum, A. (1990), 'Prospect theory and the risk-return association : An empirical examination in 85 industries', *Journal of Economic Behavior & Organization* **14**(2), 187–203.

Fischer, C. (2003), Output-based allocation of environmental policy revenues and imperfect competition, Discussion Papers dp-02-60, Resources For the Future.

Fischer, C. & Fox, A. K. (2007), 'Output-based allocation of emissions permits for mitigating tax and trade interactions', *Land Economics* **83**(4), 575–599.

Goulder, L. H., Parry, I. W. H., Williams III, R. C. & Burtraw, D. (1999), 'The cost-effectiveness of alternative instruments for environmental protection in a second-best setting', *Journal of Public Economics* **72**(3), 329–360.

Hanushek, E. A. (1974), 'Efficient estimators for regressing regression coefficients', *The American Statistician* **28**(2), 66–67.

Harvey, S. & Hogan, W. (2002), Market power and market simulations, Mimeo, Center for Business and Government, John F. Kennedy School of Government, Harvard University.

Hirshleifer, D. (2001), 'Investor psychology and asset pricing', *Journal of Finance* **56**(4), 1533–1597.

Jensen, J. & Rasmussen, T. N. (2000), 'Allocation of co2 emissions permits: A general equilibrium analysis of policy instruments', *Journal of Environmental Economics and Management* **40**(2), 111–136.

Maddala, G. (1983), *Limited-Dependent and Qualitative Variables in Econometrics*, Cambridge University Press, New York.

Montgomery, D. W. (1972), 'Markets in licenses and efficient pollution control programs', *Journal of Economic Theory* (3), 395–418.

National Commission on Energy Policy (2007), Allocating allowances in a greenhouse gas trading system, Staff paper, NCEP.

Neuhoff, K., Martinez, K. K. & Sato, M. (2006), 'Allocation, incentives and distortions: the impact of eu ets emissions allowance allocations to the electricity sector', *Climate Policy* **6**(1), 73–91.

Quirion, P. & Demailly, D. (2006), 'CO2 abatement, competitiveness and leakage in the European cement industry under the EU ETS: grandfathering versus output-based allocation', *Climate Policy* **6**, 93–113.

Sarin, R. K. & Weber, M. (1993), 'Effects of ambiguity in market experiments', *Manage. Sci.* **39**(5), 602–615.

Sijm, J., Neuhoff, K. & Chen, Y. (2006), Co2 cost pass through and windfall profits in the power sector, Cambridge Working Papers in Economics 0639, Faculty of Economics, University of Cambridge.

Sterner, T. & Muller, A. (2008), 'Output and abatement effects of allocation readjustment in permit trade', *Climatic Change* **86**(1-2), 33–49.

U.S. Environmental Protection Agency (2009), The United States Environmental Protection Agency's Preliminary Analysis of the Waxman-Markey Discussion Draft in the 111th Congress, The American Clean Energy and Security Act of 2009, Technical report, http://www.epa.gov/climatechange/economics/economicanalyses.html.

Wolfram, C. D. (1999), 'Measuring duopoly power in the british electricity spot market', *American Economic Review* **89**(4), 805–826.

Zachmann, G. & von Hirschhausen, C. (2008), 'First evidence of asymmetric cost pass-through of eu emissions allowances: Examining wholesale electricity prices in germany', *Economics Letters* **99**(3), 465 – 469.

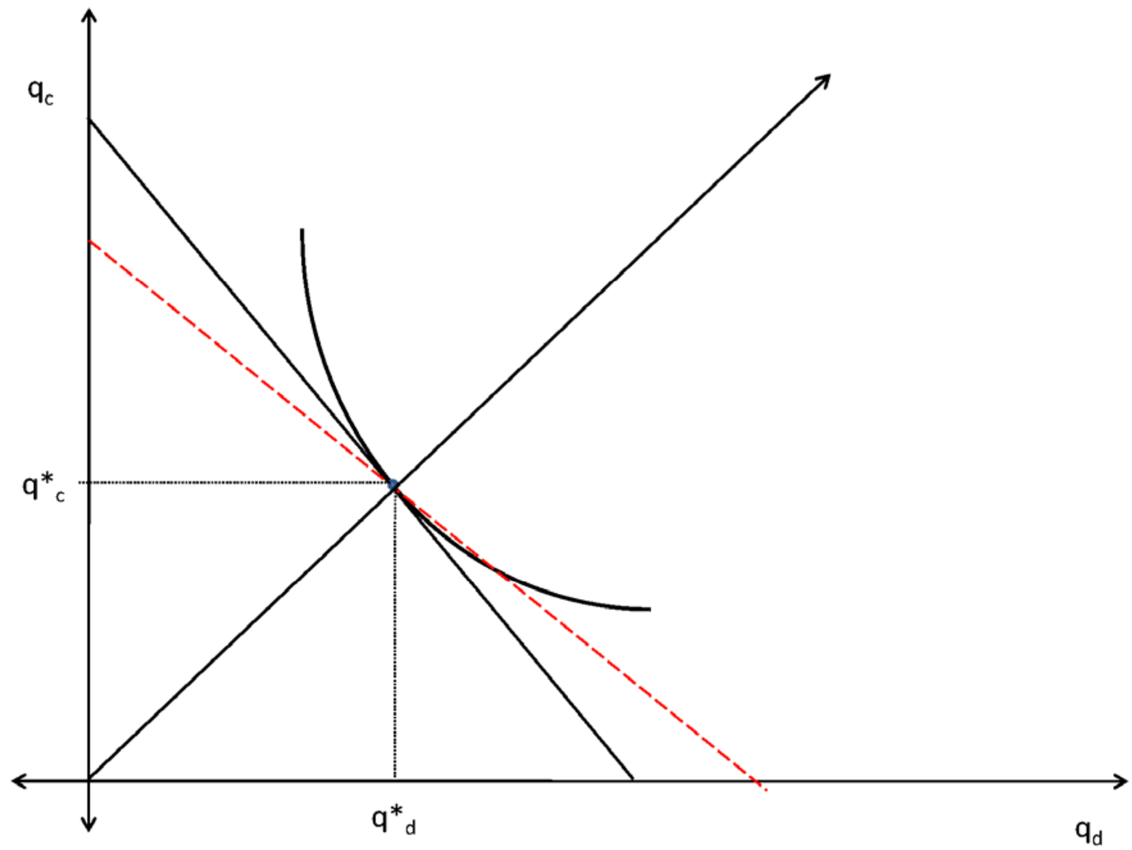

Figure 1 : Emissions constrained social welfare maximizing outcome

Notes : The downward sloping solid line represents the emissions constraint. The socially optimal allocation of production occurs at the point where the emissions constraint is just tangent to a level set of the economic surplus function. This point is intersected by the vector from the origin. The broken line, with a slope of -1, connects all points that correspond to an aggregate output quantity equal to that associated with the optimum outcome. Points lying on the emissions constraint above (below) the optimal point are associated with more (less) consumption and more (less) supply side abatement than is consistent with welfare maximization.

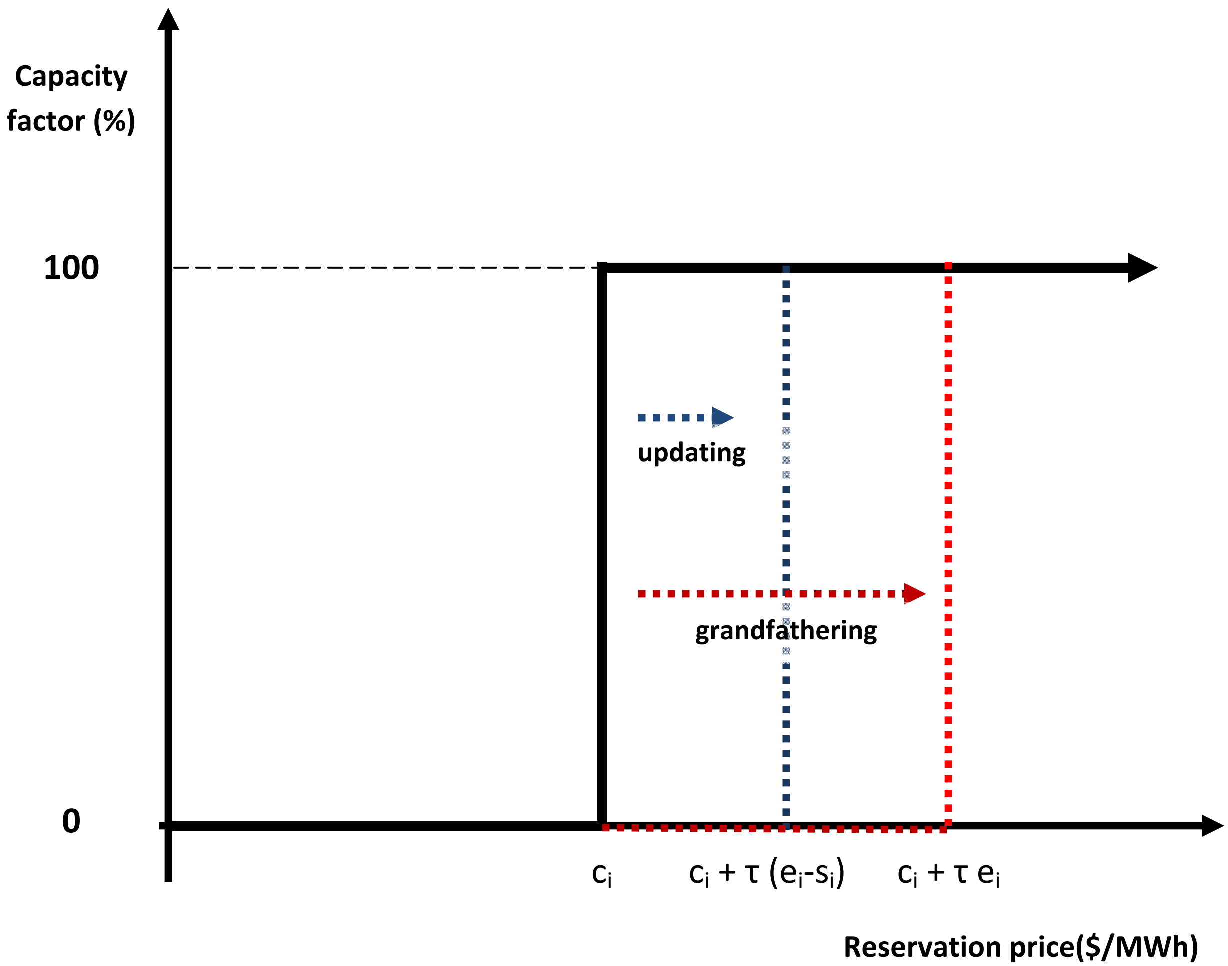

Figure 2 : Supply response implied by the static model

Notes : The vertical axis measures capacity factor. The static model predicts that an electricity producer will be inactive (at full capacity) when the price falls below (exceeds) marginal operating costs. In ozone season, the model predicts that the reservation price of a producer operating in a grandfathering (or auctioning) regime will shift right by an amount equal to the product of the permit price and the firm's emissions. For firms operating in contingent updating regimes, the shift in the reservation price is equal to the product of the permit price and NOx emissions rate less the monetized present value of the implicit subsidy.

τ

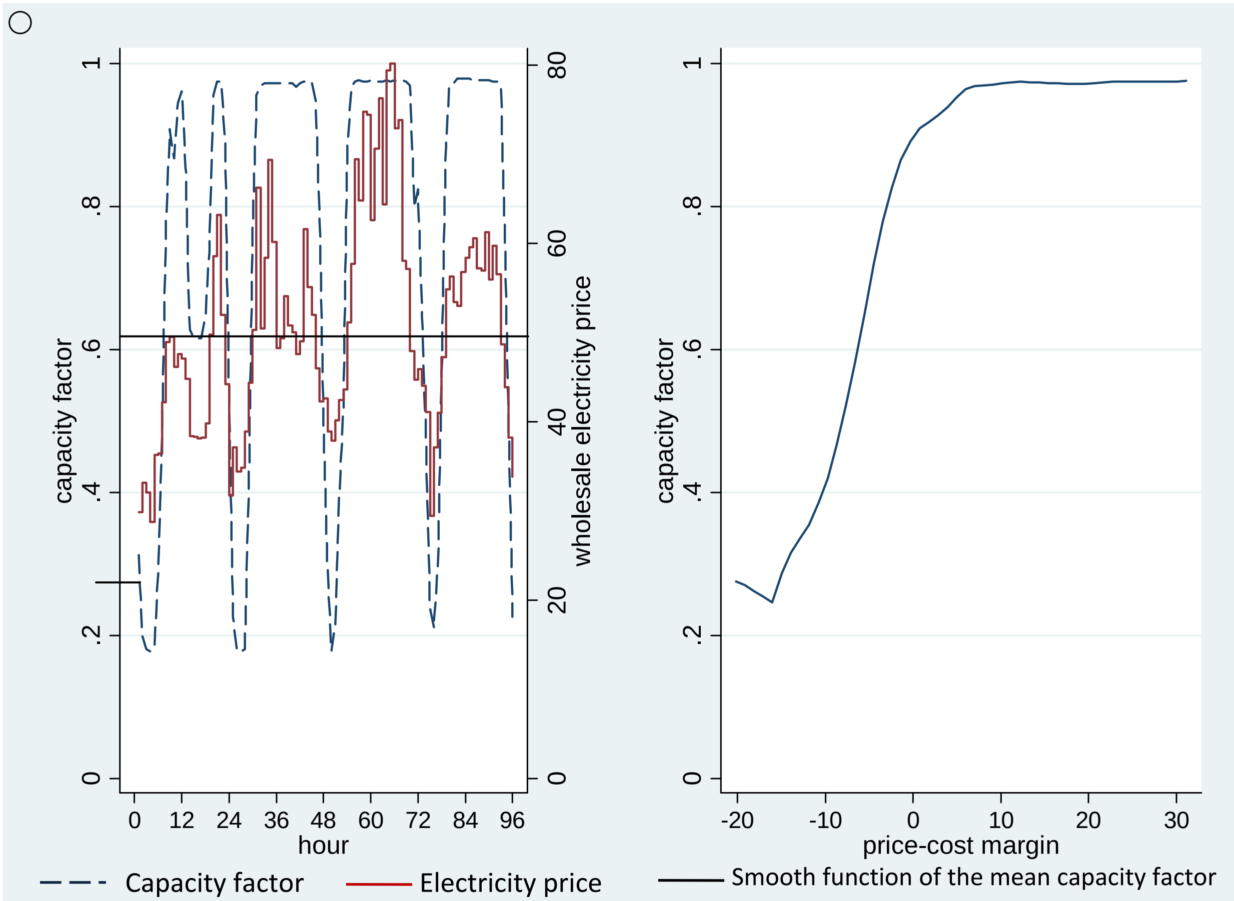

Figure 3 : Hourly Production Decisions at a Representative Unit

Notes : Hourly production decisions at a single unit (measured as capacity factor) and the corresponding hourly wholesale electricity price over a four day period are plotted in the left panel. The horizontal line represents the theoretical reservation price during these off-season hours (i.e. the unit's constant marginal operating cost). This cost of $49/MWh is estimated using the fuel input prices that prevailed in this four day period, the unit-specific heat rate, and other variable (non-fuel) operations costs. The thin black line in the right panel plots these same data in capacity factor, price-cost margin (i.e. the hourly wholesale electricity price less the marginal operating costs incurred) space. This is a mean smooth of capacity factor on price-cost margins. The thick black line represents the on-off production protocol implied by the benchmark model of a profit maximizing, price taking producer. Comparing these two functions helps to illustrate how observed production decisions deviate systematically from the predictions of the simple, static model. Finally, the thick red linerepresent the expected impact of the NBP on these short-run supply functions. In theory, the unit's reservation price should increase by an amount equal to the net compliance costs per MWh.

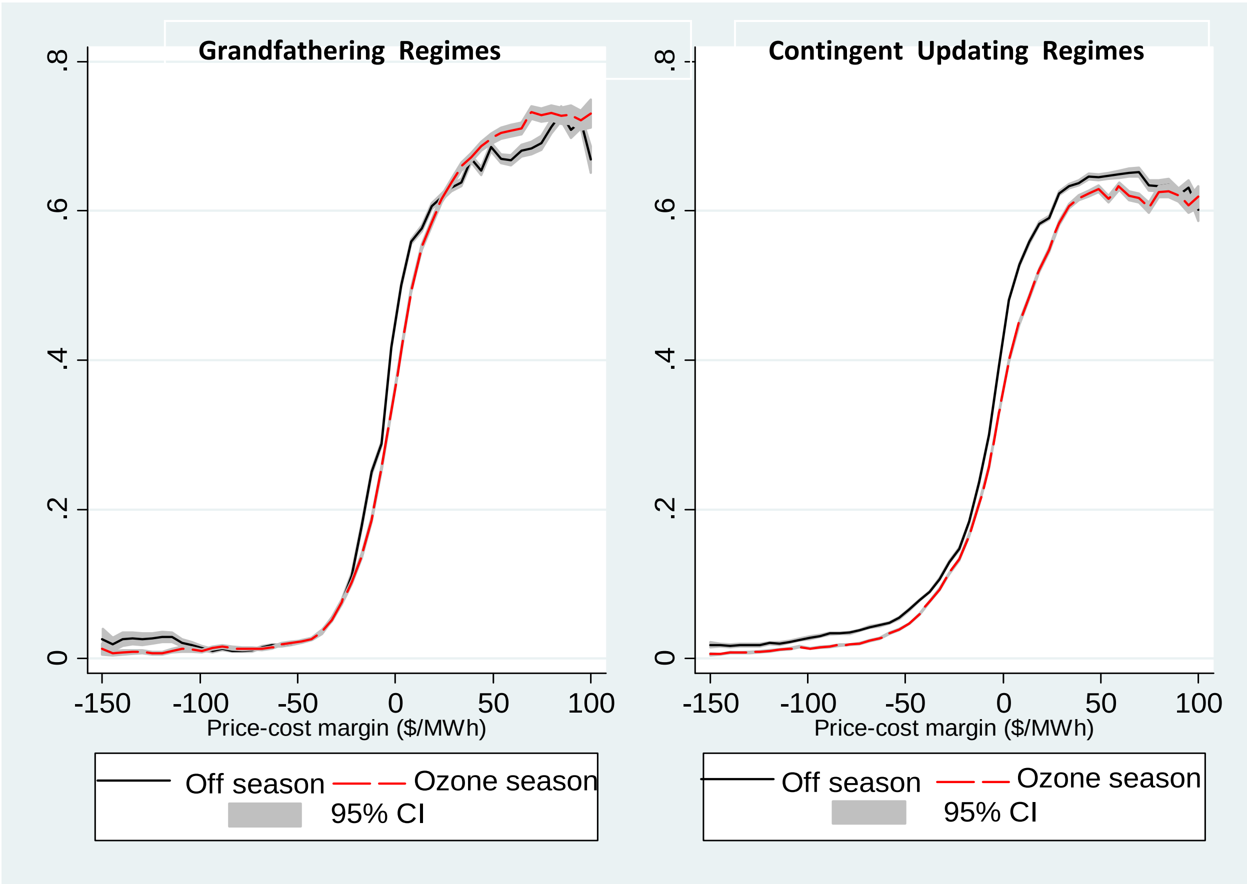

Figure 4 : Electricity Supply Decisions in Grandfathering and Contingent Allocation Updating Regimes

Notes : These figures plot a local mean smooth of hourly, unit-level capacity factors on hourly, unit-specific price-cost margins, where costs include fuel costs and other variable operating costs but exclude any costs or implicit subsidies associated with the NOx Budget Program. The shaded regions represent 95 percent confidence intervals. These graphs are generated using hourly from almost 600 electricity generating units over the study time period (2003-2006), excluding the top and bottom five percentiles of observations. I use an Epanechnikov weight function and rule-of-thumb bandwidth estimator. The right panel summarizes the production decisions at 73 units operating in states where NOx permits are grandfathered. The left panel summarizes production decisions at 515 units operating in states where permit allocations are periodically updated based on lagged input or output choices.

Table 1 : Permit allocation regime chosen by New York, New England, and Mid-Atlantic states

Electricity market	State	Annual state NOx budget for Electricity generating units (tons NOx)	Chosen permit allocation regime
New England	CT	4,253	Output-based updating
	MA	12,861	Output-based updating
	ME	N/A	N/A
	NH	N/A	N/A
	RI	936	Grandfathering
	VT	N/A	N/A
New York	NY	30,405	Input-based updating
PJM	DC	233	Grandfathering
	DE	4,463	Grandfathering
	MD	14,520	Grandfathering
	NJ	8,200	Output-based updating
	PA	47,244	Input-based updating
	VA	17,091	Input-based updating

Notes: Annual, state-level NOx budgets do not change over the study period. Among states that chose contingent updating, implementation details vary considerably. For example, whereas some states update annually, others update in three or four year blocks.

Table 2 : Operating summary statistics by permit allocation regime : CENSUS

Allocation regime	# Units	Summer Capacity (MW)	Off-season capacity factor	Heat rate* (btu/kWh)	Ozone season NOx rate* (lbs NOx/MWh)	Ramp rate (percent)
Input-based updating	353	155 (206)	19% (26%)	11,384 (2,444)	2.01 (2.04)	50.85 (29.70)
Output-based updating	140	134 (176)	10% (18%)	12,857 (3,090)	2.82 (4.54)	54.19 (26.49)
Grandfathering	69	200 (182)	21% (26%)	12,080 (3,447)	3.07 (3.60)	42.71 (22.26)
Exempt	18	181 (234)	23% (29%)	11,592 (5,388)	1.57 (4.67)	46.29 (32.03)

Notes: Standard deviations in parentheses. Summary statistics are generated using data from 610 fossil fuel-fired electricity generating units supplying the New York, New England, or PJM markets during the study period (2003-2006). Self generating and co-generating units are excluded from the sample.
* Emissions rate and heat rate summary statistics are weighted by installed capacity.

Table 3 : NOx Allowance Prices 2003-2006 (Nominal $/ton)

Permit vintage	Transaction year			
	2003	**2004**	**2005***	**2006***
2003	**$3682**	$1906	-	-
2004	$3163	**$2250**	$2180	$1507
2005	$2204	$3432	**$2771**	$1507
2006		$2951	$3018	**$1842**
2007		$2665	$2705	$1750
2008		$2705	$2299	$1570
2009		$2314	$2232	$1518

Notes: This table reports average annual permit prices by NOx permit vintage. Contemporaneous permit prices appear in bold. The asterisk denotes years in which the progressive flow control (PFC) constraint was binding. The PFC ratio was 0.25 and 0.27 in 2005 and 2006, respectively; banked permits are traded at a discount in these years.

Table 4 : Estimated NBP compliance costs and production incentives by allocation regime

Allocation regime	(1) NOx permit costs per MWh generated	(2) (1) as a percentage of off-season variable operating costs	(3) Future permits allocated (tons) per MWh generated	(4) Estimated net compliance cost per MWh*
Input-based updating	\$4.89 (\$4.69)	6.4% (6.0%)	0.001 (0.001)	\$1.40 (\$4.69)
Output-based updating	\$7.46 (\$8.24)	6.5% (5.9%)	0.002 (0.001)	\$1.45 (\$8.15)
Grandfathering	\$6.49 (\$9.03)	8.1% (6.8%)	0	\$6.49 (\$4.01)
Exempt	\$0	--	0	\$0

Notes: Standard deviations in parentheses. Summary statistics are generated using data from 580 fossil fuel-fired electricity generating units supplying the New York, New England, or PJM markets during the study period (2003-2006). Self generating and co-generating units are excluded from the sample. Averages across unit-years are reported; standard deviations are in parentheses.

* To calculate net compliance costs, future permits allocated per unit of output are valued using futures permit prices. This value is then subtracted from the explicit compliance cost per MWh (i.e. the product of the unit-specific emissions rate and the spot NOx permit price).

Table 5 : Regional wholesale electricity prices and electricity demand by season

Regional electricity market	Hourly electricity demand (MW)		Wholesale electricity price ($/MWh)	
	Off-season	Ozone season	Off-season	Ozone season
New England (NEPOOL)	14,603 (2,513)	15,344 (3,416)	$60.88 ($29.34)	$57.98 ($34.88)
New York (NYISO)	17,547 (2,768)	19,151 (3,936)	$68.00 ($96.97)	$70.74 ($78.17)
Mid-Atlantic (PJM)	29,214 (8,787)	31,694 (9,957)	$52.03 ($33.84)	$55.41 ($43.34)

Notes: Standard deviations in parentheses. This table summarizes observed prices and load levels over the 32,748 hours in the data. Both prices and load levels are similarly distributed across seasons.

Table 6 : First stage point estimates

Covariate	H=6	H=2	H=8
Unit-level constant	-3.04	-2.77	-3.12
Marginal operating cost ($/MWh)	-0.01	-0.01	-0.01
Local wholesale marginal price ($/MWh)	0.08	0.005	-0.008
1 hour Forward price ($/MWh)	-0.02		0.03
2 hour forward price ($/MWh)	0.04		-0.02
3 hour forward price ($/MWh)	-0.03		0.02
4 hour forward price ($/MWh)	0.03		-0.003
5 hour forward price ($/MWh)	-0.03		-0.000
6 hour forward price ($/MWh)	0.02		0.000
7 hour forward price ($/MWh)			0.000
8 hour forward price ($/MWh)			0.000
NOx price * Ozone indicator ($/ton)	-0.0001	-0.0002	-0.001

Notes: The unit of analysis is a unit-hour. The dependent variable is the binary participation indicator. Each unit-level regression includes: a unit-level fixed effect, marginal operating costs, contemporaneous wholesale price , hourly forward prices, the NOx permit price interacted with the ozone season indicator and year fixed effects. Real time and day ahead electricity prices are instrumented for using hourly demand forecasts.

Table 7 : Second stage estimation results

Specification	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Constant	0.000 (0.000)	-0.003 (0.002)	-0.004*** (0.001)	-0.002 (0.001)	-0.002 (0.002)	-0.002 (0.002)	0.002 (0.002)	-0.005* (0.001)
NOx rate (tons/MWh)	-2.21*** (0.80)	-2.44*** (0.82)	-3.04*** (0.44)	-2.05*** (0.49)	-2.87*** (0.54)	-2.97*** (0.65)	-2.73*** (0.52)	-3.09*** (0.45)
Estimated subsidy (tons/MWh)		1.69 (1.23)	1.92*** (0.70)	1.45* (0.76)	1.48* (0.84)	1.47 (0.99)	1.69** (0.71)	1.65** (0.75)
NOx rate * capacity			-0.008** (0.002)	-0.002 (0.003)	-0.003 (0.004)	-0.014*** (0.01)	-0.007* (0.004)	-0.008** (0.003)
Estimated subsidy * capacity			0.001* (0.004)	0.007* (0.004)	0.004 (0.004)	0.010*** (0.005)	-0.003 (0.005)	-0.001 (0.003)
Ramp rate							-0.003** (0.001)	
Capacity							0.00 0.00	
NE								-0.001 (0.002)
NY								0.003* (0.0021)
R^2	0.09	0.10	0.11	0.08	0.10	0.06	0.12	0.12
N	550	550	550	553	550	553	550	550
Assumed time horizon (hours)	6	6	6	6	2	8	6	6
First stage IV?	Y	Y	Y	N	Y	Y	Y	

Notes: The unit of analysis is an electricity generating unit. The dependent variable is the coefficient on the interaction between the NOx permit price and ozone season indicator from the first stage of the estimation. First stage probit equations also include a constant, unit-specific marginal operating cost, year indicators, and instruments for current and future wholesale electricity prices (unless otherwise indicated). These first stage equations are estimated under different assumptions about the relevant time horizon (measured in hours). Emissions rates (demeaned) and estimated subsidies are measured in tons of NOx per MWH. Production capacity (also demeaned) is measured in MW. FGLS standard errors (in parentheses) are clustered at the facility level. See text for details

* Statistically significant at the 10 percent level.
** Statistically significant at the 5 percent level.
*** Statistically significant at the 1 percent level.